
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

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 2: The Future Role of AI in Warfare

Manchester

ANITA ANAND: Welcome to the second Reith Lecture with one of the world's leading authorities on artificial intelligence, Professor Stuart Russell, from the University of California at Berkeley. Now, today we're in the north of England, it's not quite California, but we are at the University of Manchester in the magnificent neogothic splendour of the Whitworth Hall, and we're going to hear Stuart's ideas on the future role of AI in Warfare.

Now this, as you know, has been a subject for a very long time that has taken up the vivid imagination of those who make films and write fiction. How will the wars of the future be fought? What might that mean for us? Will AI reduce collateral damage and civilian casualties, or will AI kill on a scale not seen since Hiroshima and Nagasaki?

I think it's time we hear what our lecturer thinks. Will you please welcome the BBC's 2021 Reith Lecturer, Professor Stuart Russell.

(AUDIENCE APPLAUSE)

STUART RUSSELL: The story this evening begins on the 20th of February 2013, when a rather puzzling email arrived from Human Rights Watch, or HRW. I had been a member of the HRW Northern California committee for some time. HRW is an incredible organisation. For more than 40 years it has investigated atrocities around the world, atrocities committed by humans. Now, HRW was asking me to support a new campaign to ban "killer robots." The letter raised the possibility of children playing with toy guns being accidentally targeted by the killer robots. It stated that robots would not be restrained by "human compassion" which can provide an important check on the killing of civilians. So now it's "Humans good, robots bad"?

Apparently, I recovered well enough from my initial confusion to reply, two hours later, saying I'd be happy to help. I thought perhaps we could start with a professional code of conduct for computer scientists, something like, "Do not design algorithms that can decide to kill humans," but we would need clearer arguments to convince people to sign on.

My goal today is to explain those arguments and how they have evolved but let me begin with some caveats. First, I am not talking about all uses of AI in military applications. Some uses, such as the better detection of surprise attacks, could actually be beneficial.

Second, this is not about the general morality of defence research. I think we would all prefer to have no wars, but if your taxes are paying someone to die in your defence, it's hardly a moral position to refuse to help protect them.

Finally, I'm not talking about drones in the sense of aircraft that are remotely piloted by humans. Everyone in arms control knows that the US is very sensitive about not sweeping their drones into this discussion, so now we reflexively say, "We're NOT talking about human-piloted drones," as I just did.

The technical term for my subject today is lethal autonomous weapons systems, which means, according to the United Nations, “weapons that locate, select, and engage human targets without human supervision.” The word ‘engage’ here is a euphemism for ‘kill’.

Right now, I suspect you’re imagining a rampaging Terminator robot, and if you weren’t, you are now. I’ve tried to convince journalists to stop using this image for every single article about autonomous weapons and I’ve failed miserably. I suspect the movie franchise is paying them.

This Terminator picture is wrong for so many reasons. First of all, the Terminators fire a lot of bullets that miss their targets. Why do they do that?

Secondly, it makes people think that autonomous weapons are science fiction. They are not. You can buy them today. They are advertised on the web.

Third, it makes people think that the problem is SkyNet, the global software system that controls the terminators. It becomes conscious, it hates humans, and it tries to kill us all. I went to a meeting where the US Deputy Secretary of Defence said, “We have listened carefully to these arguments and my experts have assured me that there is no risk of accidentally creating SkyNet.” He was deadly serious.

Let me assure you of the same thing. SkyNet never was the problem. If you want a better picture from science fiction, think about the TV series *Black Mirror*, and specifically the robot bees from the episode *Hated in the Nation*. They aren’t conscious. They don’t hate people. They are precisely programmed by one person to hunt 387,036 specific humans, burrow into their brains, and kill them.

If you’ve seen that episode, you’re probably wondering, “Why is he even talking about this? Surely no one in their right mind is going to produce weapons like that!” I wish that were true. Or perhaps it is true, and a lot of people aren’t in their right minds.

Let’s try to understand how we got to where we are today, with lethal autonomous weapons advertised on the web.

Weapons are governed in part by international humanitarian law, which includes the Geneva Conventions, in particular, The Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. For some reason this is known by just three of its initials, CCW.

One of the main rules of international humanitarian law is the principle of distinction: you cannot attack civilians, and, by extension, you cannot use weapons that are by nature indiscriminate. A UN report in 2012 warned that autonomous weapons were coming and might be indiscriminate, they might accidentally target civilians, especially in the "fog of war." From this warning came HRW's report, with its example of children being targeted because they're playing with toy guns.

For reasons that will become clear, I think this focus on accidental targeting was a mistake, but at that time it was the primary concern and it led to CCW's first discussion of autonomous weapons in Geneva in 2014.

I didn't go to that first meeting, but I heard that the collision of suave diplomats and complicated technological issues was not pretty. Total confusion ensued, especially around the meaning of the word "autonomous."

For example, Germany's official position was that a weapon is autonomous only if it has "the ability to learn and develop self-awareness." In other words, only SkyNet counts. China, which ostensibly supports a ban on autonomous weapons, says that as soon as weapons become capable of autonomously distinguishing civilians and soldiers, they no longer count as autonomous and so they wouldn't be banned.

In 2015, I was invited to address the CCW meeting in Geneva as an AI expert. I had three jobs to do: clear up the mess with autonomy, assess the technological feasibility of autonomous weapons, and evaluate the pros and cons as best I could. It seemed to me, in my naïveté, that this was a chance to steer the discussion in a sensible direction.

Explaining autonomy didn't seem that difficult. It's exactly the same kind of autonomy that we give to chess programs. Although we write the chess

program, we do not decide what moves to make. We press the start button, and the chess program makes the decisions. Very quickly it will get into board positions no one has ever seen before, and it will decide, based on what it sees and its own complex and opaque calculations, where to move the pieces and which enemy pieces to kill or, should I say, “engage.”

That’s exactly the UN definition: weapons that locate, select, and engage human targets without human supervision. There’s no mystery, no evil intent, no self-awareness; just complex calculations that depend on what the machine’s camera sees - that is, on information that is not available to the human operator.

The next question I needed to explain to the CCW was feasibility. Could these weapons be built with current technologies? Let’s look at the components one by one.

First, there must be a mobile platform. Even at the time of my Geneva lecture in 2015 there were already many options: quadcopters ranging from 3 centimetres to 1 metre in size; fixed-wing aircraft ranging from hobby-sized package delivery planes to full-sized missile-carrying drones and “autonomy-ready” supersonic fighters like the BAE Taranis; self-driving cars, trucks, tanks; prototype submarines and destroyers; and even, if you must, skeletal humanoid robots. There were demonstrations of quadcopters catching balls in mid-air, flying sideways through vertical slots at high speed; even large formations of them filing through narrow windows and re-forming inside buildings. Nowadays perfectly coordinated aerobatic swarms of over 3000 quadcopters are routine.

Next, the machine must be able to perceive its environment. In 2015, the algorithms already deployed on self-driving cars could track moving objects in video, including human beings and other vehicles. Autonomous robots could already explore and build a detailed map of a city neighbourhood or the inside of a building. I found a creepy video of a quadcopter going into a house, exploring and mapping the ground floor, and then heading upstairs to search the bedroom.

Then there’s the ability to make tactical decisions. These might resemble the ones demonstrated by AI systems in multiplayer video games or in self-driving cars, but while a self-driving car can never make a serious mistake, a

lethal weapon that works 90 per cent of the time is just fine, so the weapon's problem is actually much easier.

Finally, there is the killing or, should I say, the engaging part. I couldn't remember having taken any courses in "engaging," so I had to educate myself.

Some weapons were already available on remotely piloted drones, including vision-guided missiles, gyro-stabilized machine guns, and kamikaze weapons like the 50-pound explosive in the nose of Israel's Harpy loitering missile.

I also spent a few unpleasant days learning about shaped charges and explosively formed penetrators. I was particularly struck by a demonstration that a shaped charge the size of a fizzy drink can could penetrate several feet of steel plate. In a minute we'll see why this is relevant.

So as far as feasibility was concerned, in 2015 all the component technologies for autonomous weapons already existed and it would not be too hard to put them together. Strangely, the arms control community, including HRW and a group of 20 Nobel peace prize-winners, kept saying that these weapons "could be developed within 20 to 30 years." On the other hand, my robotics colleagues said, "Eighteen months, tops". Britain's Ministry of Defence said "probably feasible now" for some scenarios.

Finally, the pros and cons: should we develop and deploy autonomous weapons, or should we ban them? Before I stuck my neck out, I needed to understand all the arguments.

One potential benefit of autonomy is that wars fought between robot armies might avoid human casualties altogether. But if that were true, we could also settle wars by playing tiddlywinks. In the real world, wars end when the level of death and destruction becomes untenable for one side, or for both.

At the CCW, the American and British delegations claim that autonomous weapons can "reduce civilian casualties due to more precise targeting." This is an extension of the argument they make for remotely piloted drones. The extension hinges on two assumptions.

The first is that AI systems will be better than humans at recognizing legitimate targets. This claim was probably false in 2015 but it was also a moving target: I could not say that it would never be true.

The second, and usually unstated assumption, is that autonomous weapons will be used in essentially the same scenarios as human-controlled weapons, including human-controlled weapons such as rifles and tanks that are actually attached to humans. This seems to me unequivocally false. And if autonomous weapons are used more often, by different parties, against different targets, with different goals, and in less clear-cut settings, then any putative advantage in distinguishing civilians and soldiers is irrelevant. For this reason, I think the emphasis on this question has been misguided.

So much for the pros of autonomous weapons.

As for the cons: well, there are obvious practical objections, such as the fact that autonomous weapons might be subject to cyber-infiltration, causing them to turn against their owners once a war started.

Accidental escalation of hostilities is also a real concern: if defence systems overreact, a false alarm leads to a real retaliation, and things escalate quickly into a real war.

Both cyber-infiltration and escalation are already taken seriously by military planners, but they seem to be plunging ahead regardless.

Campaigners have also raised legal arguments, such as the supposed “accountability gap” that arises when AI systems commit atrocities. But my lawyer friends assured me that there was no new gap here between criminal intent and criminal negligence.

International humanitarian law also includes an explicitly moral element called the Martens Clause, which says that “in cases not covered by the law in force, the human person remains under the protection of the principles of humanity and the dictates of the public conscience.”

One can see echoes of this principle in various public statements: for example, Antonio Guterres, the UN Secretary General, tweeted, “Machines with the power and discretion to take human lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law.”

And in a surprise move to open a debate at the World Economic Forum, in which I participated, Sir Roger Carr, Chairman of BAE Systems, one of the largest defence manufacturers in the world, admitted that delegating kill decisions to machines was “fundamentally wrong” and pledged that his company would never allow it.

Even Paul Selva, Vice-Chairman of the Joint Chiefs of Staff in the United States, told Congress, “I don't think it's reasonable for us to put robots in charge of whether or not we take a human life.”

I’ve had many meetings with high-level military officers from several countries, and I am struck by how seriously they take their responsibility for life and death and by their sense of honour as soldiers. And of course, they understand that they would one day be on the receiving end of attacks by autonomous weapons, which would make the battlefield essentially uninhabitable for humans.

I didn’t believe, however, that arguments based on morality and honour alone would sway the governments that make decisions in international affairs, especially when they distrust the morality and honour of all the other governments.

The final question I explored while preparing for the CCW meeting was the future evolution of autonomous weapons. What kinds of weapons would AI enable, and how would they be used?

It seemed to me that AI would enable a lethal unit to be far smaller, cheaper and more agile than a tank, or an attack helicopter, or even a soldier carrying a gun. A lethal AI-powered quadcopter could be as small as a tin of shoe polish. And this is where the shaped charges and explosively formed penetrators come in: about 3 grams of explosive are enough to kill a person at close range.

A weapon like this could be mass-produced very cheaply. A regular shipping container could hold a million lethal weapons, and because, by definition, no human supervision is required for each weapon, they can all be sent to do their work at once. And if we know anything about computers, it's this: if they can do something once, they can do it a million times.

The inevitable endpoint is that autonomous weapons become cheap, selective weapons of mass destruction. Clearly, this would be a disaster for international security. Rather than appealing to morality or honour, I hope to appeal to nations' self-interest.

After my presentation, I found myself in the unusual position of being extremely popular with the ambassadors from Cuba, Pakistan and Venezuela, but not with the Americans and British, whose delegations sat there, stony-faced.

Their disgruntlement came, I suspect, from what they saw as Realpolitik: the overriding need to maintain military superiority over potential enemies who would develop AI weapons. I think they are missing the point, a point, in fact, that they have agreed to previously.

In 1966, a coalition of American biologists and chemists wrote to President Johnson explaining that biological weapons, which the US was developing, would, once perfected, become cheap, widespread weapons of mass destruction that would ultimately reduce American security. Eventually, Henry Kissinger convinced Johnson's successor, Richard Nixon, that the argument was valid. Nixon unilaterally renounced biological weapons and Britain drafted an international treaty to ban them.

The same argument applies to anti-personnel autonomous weapons. They could wipe out, say, all males between 12 and 60 in a city, or all visibly Jewish citizens in Israel. Unlike nuclear weapons, they leave no radioactive crater, and they keep all the valuable physical assets intact. And unlike nuclear weapons, they are scalable. Conflicts can escalate smoothly from 10, to a thousand, to a hundred thousand casualties with no identifiable calamitous threshold being crossed.

Soon after that Geneva meeting, the AI community launched an open letter calling for a ban; tens of thousands of researchers signed, including almost the entire leadership of the field. Over 2,500 media articles appeared in 50 countries. Even the Financial Times, not exactly the tree-huggers' house journal, supported the ban, calling autonomous weapons "a nightmare we have no cause to invent."

Progress! Or so I thought. Still, confusion reigned. The quote about SkyNet from the US Deputy Secretary of Defence came from a meeting at West Point in July 2016, a year after the FT editorial. Evidently, the message was still not getting through. We needed something more than earnest, well-argued articles and PowerPoint; we needed a movie, making our argument in graphic detail.

We found some brilliant writers and filmmakers at Space Digital here in Manchester and we made a film called Slaughterbots. It had two storylines: one, a sales pitch by the CEO of an arms manufacturer, demonstrating the tiny quadcopter and its use in targeted mass attacks; the other, a series of unattributed atrocities including, I'm sorry to say, the assassination of hundreds of students at the University of Edinburgh, where I'll be speaking next.

The reactions elsewhere were mostly positive. The film had about 75 million views on the web, and I'm pleased to say that CNN called it "the most nightmarish, dystopian film of 2017." Many of my AI colleagues thought the CEO's presentation was real, not fictional, which tells you something about where the technology is.

At the CCW, on the other hand, the Russian ambassador retorted, "Why are we discussing science fiction? Such weapons cannot exist for another 25 or 30 years!"

Three weeks later, a government-owned manufacturer in Turkey announced the Kargu drone, advertising its capabilities for "anti-personnel autonomous hits" with "targets selected on images and face recognition." Just like those robot bees. The website has since been altered.

According to the UN, Kargus were used last year in the Libya conflict, despite a strict arms embargo, to autonomously "hunt down" members of one of

the factions. The Kargu is the size of a dinner plate and carries a kilogram of explosive so it can destroy vehicles and attack buildings as well as people. It is likely that many similar weapons, both larger and smaller, are under development.

At the moment we find ourselves at an unstable impasse, unstable because the technology is accelerating.

On the one side, we have about 30 countries who are in favour of a ban, as well as the EU parliament, the United Nations, the non-aligned movement, hundreds of civil society organizations, and according to recent polls, the great majority of the public all over the world.

On the other side, we have the American and Russian governments, supported to some extent by Britain, Israel and Australia, arguing that a ban is unnecessary and that lethal weapons are good for you.

In the CCW, a potentially meaningless agreement seems to be in the works, banning only weapons that operate "completely outside any framework of human control," which seems to mean weapons that wake up one morning and decide to start a war by themselves. In other words, we're back to SkyNet.

Two years ago, before COVID, a small group of experts met in a house in Boston, covering the entire spectrum from advocates to opponents of autonomous weapons. After two days of arguing we reached a compromise solution: a ban that would require a minimum weight and explosive payload so as to rule out small antipersonnel weapons.

There's an interesting precedent called the St. Petersburg Declaration of 1868. Its origins seem almost quaint today: a Russian engineer had invented a musket ball that exploded inside the body, and the Imperial Court was afraid that this would be viewed as dishonourable and ungentlemanly. So, they convened a meeting, and the Declaration banned exploding ordnance below 400 grammes. To a good approximation this still holds today.

A similar ban on small anti-personnel weapons would eliminate swarms as weapons of mass destruction. It would allow the major powers to keep their big-

boy toys: submarines, tanks, fighter aircraft. The International Committee of the Red Cross, which has statutory responsibility for the Geneva Conventions, supports this solution.

Progress! Or so I thought. As the Indian ambassador kindly reminded me in Geneva, I don't understand the first thing about diplomacy.

Diplomats from both the UK and Russia express grave concern that banning autonomous weapons would seriously restrict civilian AI research. Funnily enough, I've not heard this concern among civilian AI researchers. Biology and chemistry seem to be humming along, despite bans on biological and chemical weapons. And AI researchers do not want videos of robots hunting and killing children to be the most salient example of AI in the public's mind. They might even be more comfortable contributing to defence-related AI research if autonomous weapons were off the table.

The last line of resistance of the diplomats is verification and enforcement. The US, in particular, won't sign a treaty that allows others to cheat. Here, I agree with the diplomats. I've spent a good part of the last decade improving the verification arm of the Nuclear-Test-Ban Treaty, and I would be happy to do the same now for a ban on autonomous weapons. The AI community already has good ideas, some of them borrowed from the Chemical Weapons Convention, and we're ready to start work tomorrow morning.

This lecture will be broadcast just before Sixth Review Conference of the CCW in Geneva. Let me say this to the diplomats, and to their political masters, with all due respect: There are 8 billion people wondering why you cannot give them some protection against being hunted down and killed by robots. If the technical issues are too complicated, your children can probably explain them.

Thank you.

(AUDIENCE APPLAUSE)

ANITA ANAND: Stuart, thank you very much indeed, and your liberal use of Slaughterbots, killer robots – I mean, sleep well everybody who's heard this lecture. You did talk about your desire to see something like the comprehensive

Test-Ban Treaty, for example, to do with AI, but may I put it to you that we have a comprehensive Test-Ban Treaty, and we have nuclear proliferation. Isn't the genie just out of the bottle now?

STUART RUSSELL: Interestingly, the comprehensive Test-Ban Treaty, despite being proposed by the United States in 1958, was not ratified by the United States in 1997 and, as a result, it hasn't entered into force, so it doesn't actually exist as a treaty. However, all the countries who have signed it have refrained from nuclear testing.

ANITA ANAND: It's like reliving that experience of Alfred Nobel looking upon his creation and despairing. It exists. You've said already, it exists. How can you uninvent what is already out there?

STUART RUSSELL: Well, I think we've, to a large extent, uninvented chemical weapons quite well. They were used in huge quantities in the First World War. I believe they killed about 200,000 people during the war, and then another hundred thousand people died slow, agonising deaths in the decades that followed, and now we have very, very few casualties. They were not really used in World War II and then the Chemical Weapons Convention strengthened the constraints, added clauses for verification and enforcement, which I think are very important, and although there were casualties in Syria, and Syria is not one of the signatories, they have pretty much disappeared from military planning, and they have not been a significant factor. So that's good.

Biological weapons also, they were apparently used with some success during World War II on the Russian Front and even though the Russians cheated because the Biological Weapons Convention doesn't have a verification clause, even though the Russians cheated, they have not been used to kill large numbers of people in war. And so, I think very few people would like to say, "Oh, let's get rid of the Chemical Weapons Convention and the Biological Weapons Convention."

So, arms control has been successful. Even the Land Mine Treaty, which, again, the United States and several other major powers haven't ratified, has been very effective. The vast majority of land mine manufacturers have stopped

making land mines. Tens of millions of land mines have been destroyed and many, many minefields have been removed, so it's been successful.

The ban on blinding laser weapons, again, a technology that could be quite useful in war but people decided that we really didn't need it and so we did it, we got rid of them.

ANITA ANAND: Well, there's a comprehensive answer to that question. Let me open this up now to the audience at the Whitworth Hall.

CHI CHI EKWEOZER: Hi there, I'm Chi Chi Ekweozer, I'm a tech entrepreneur and also a start-up founder. As super-intelligent AI gets more popular around the world, how worried are you about them falling into the wrong hands: terrorists, rogue states, et cetera, and how do you believe we can police this?

STUART RUSSELL: How can we police super-intelligent AI? Well, I think we're doing such a good job of policing malware and cybercrime already that I'm really not worried about this at all. No. So, actually, the current situation is disastrous with cybercrime. I mean, the software industry generates revenues of about 500 billion a year. The semi-conductor industry generates revenues of about 500 billion a year. The cybercrime industry generates revenues of about a trillion a year. So we made serious mistakes very early on, actually, in the way we developed our networks, our protocols, our software systems, without requiring authenticated identities and so on, and I think if we're going to have a chance of policing the use of increasingly capable AI systems for increasingly dastardly cybercrimes, and we're not here talking about weapons per se but just AI for automated blackmail, theft, impersonation, fraud, et cetera, et cetera, we're going to need to take some very serious steps with, I think, international agreements.

There is a Budapest Convention that about 50-something countries have signed which does allow for cross-border forensics and other sorts of things but it's not covering most of the major players in cybercrime. So, we need to make real efforts on that, otherwise the situation is going to get much worse, as you say, in future.

ANITA ANAND: Thank you very much for that question. Stuart, you were talking about this Slaughterbots movie and that a lot of people watched it and thought, well, this is real, this is already with us. I mean, it might be helpful to know what really is out there at the moment. We have with us Air Commodore David Rowland, who leads the establishment for the Defence AI Centre with the MOD. Where are we at the moment? I mean, we may not have Slaughterbots, what have we got?

AIR COMMODORE DAVID ROWLAND: Yes, Anita, thank you very much. We in defence see that AI has got some real potential benefits for us to utilise, but, of course, there's going to be some risks and some threats that we've got to think about. Right now, in defence we're looking at how can we utilise it to help us. So, for example, our intelligence staff, how can they analyse the massive amounts of data that they get to help them out, or our logistics and, indeed, even back-office type work. We utilise it in very much areas that can help us out.

I'm sure that the Professor would agree that AI is here, we are going to see future conflicts that have got AI within them. We know that our adversaries are absolutely investing in this, and it's right and incumbent upon us to make sure that we understand AI and can utilise it. I suppose the question would be is doesn't he agree that we absolutely need to utilise AI in defence and what areas should we be looking at to make sure that we do maintain a competitive advantage.

ANITA ANAND: Thank you very much. Stuart?

STUART RUSSELL: I completely agree, and we've been developing AI for all sorts of defence applications for decades and in the US, DARPA, the Defence Advance Research Projects Agency, is one of the major funders of basic research in AI and it's had major successes. For example, in Desert Storm, planning all the logistics for moving hundreds of thousands of troops and all their equipment to the Middle East in a very short period of time was achieved using AI planning systems to make sure that that would all work, and they said at the time that just that one application more than paid back the entire investment in AI over the history of DARPA.

I've had similar kinds of discussions. We had a meeting at the White House in 2016 and we talked about the same questions, the importance of AI, how it can really improve planning, logistics, reconnaissance, intelligence analysis, et cetera, et cetera, and then we got to the nitty gritty, what about the weapons of mass destruction, the huge drone swarms that could kill millions of people, and a gentleman from the National Security Council said, "But we would never make weapons like that," and the appropriate response is, "Then why not ban them? Why are you arguing against a ban on those types of weapons?" Didn't get a straight answer to that question.

ANITA ANAND: Throw it back to the Commodore. Do you want to answer it?

AIR COMMODORE DAVID ROWLAND: How do we ban something that we can't agree a definition on, so why not use those processes that are already there that are so successful in other areas?

STUART RUSSELL: I don't actually agree that we can't define them. I think that it's actually not so difficult to talk about and I have a proposal, the International Committee of the Red Cross supports this, that small anti-personnel autonomous weapons should be banned, and I don't think IHL, the International Humanitarian Law, is sufficient. It's a practical question and I'm quite confident, as you say, that the UK and probably the US, would not be using such weapons to wipe out large populations but if I was sitting in Israel, I would be quite worried that one of Israel's nearby enemies might use these types of weapons to wipe out all the Jews in Israel, and that would be terrible.

IHL, arguably, doesn't allow you to use nuclear weapons to target civilian populations but, nonetheless, we also don't sell nuclear weapons in Tesco, but these types of autonomous weapons, like the Kargu, right, Kargu is already being used in a theatre where there's an arms embargo and smaller, cheaper weapons, for example, you used to be able to buy land mines for less than $7 each. We could have small, lethal autonomous weapons costing between five and $10 each and if you allow those to be manufactured in large quantities, then it's like selling nuclear weapons in Tesco. They'll be available in the international arms market for whoever has - the price of one F35 you can buy 20 million weapons. That's not a future that makes sense to me.

ANITA ANAND: Okay, and because this is the BBC, I will say other supermarkets exist that also do not sell weapons of mass destruction. Let's take another question from the back. There was a hand over there? Thank you very much.

PHIL HORN: Hello. Phil Horn. Three-part question but very rapidly delivered. First, is war and warfare inevitable amongst humans and so is there an end to this? The second part is do you think that chemical and biological warfare died away because the next technological stage came along? And then the third part is, if that is true, what comes after AI, what's on your radar in terms of the next threat?

ANITA ANAND: Thank you very much indeed. So, is war inevitable? A little question to start off with.

STUART RUSSELL: I would like to think that it isn't inevitable but when the machine gun was first invented, one of the justifications for developing it was that it would bring an end to war because its destructive capability was so great that no-one would ever fight wars again. That didn't happen. I think some of the data suggests that the frequency of lethal conflicts is decreasing over time and so perhaps that will continue. I am an optimist by nature.

ANITA ANAND: And the second part, which I think you've partly answered already, which is chemical and biological weapons, did they just end because the next stage of destruction came along?

STUART RUSSELL: So that's an interesting question. Obviously, the Chemical Weapons Convention and the Biological Weapons Convention were happening in the nuclear era and I think there are good arguments to say that neither category of weapon is particularly precise, and sometimes not as effective as one might hope, but the Russians certainly believed that they could be extremely effective and they had experience, as I said, on the Russian Front against Germany, so they put a huge effort into this and they moved the technology a long way past where it was when the ban was first passed. So, I can't say for sure that they were outmoded weapons, but I think, actually, the stigmatisation of those weapons over time, the horrors of World War I with

chemical weapons, I think, had a significant impact on the non-use of chemical weapons in World War II. So, I think they were morally outmoded rather that sort of tactically outmoded.

ANITA ANAND: And the third part, is AI – I suppose just to paraphrase – the final frontier of weaponry or is there something else?

STUART RUSSELL: Well, I don't know if you've seen the latest James Bond film No Time to Die, but that's another weapon, and it is sort of related. I don't know how much of the plot I want to give away but there is a nanobot technology that's actually created originally by the British government as a weapon of extremely precise assassination of a particular individual, based on their DNA, and at one point in the film, I think it's Q who says, "We never intended this to turn into a weapon of mass destruction," and it's in some sense being used in the same way, to targeting entire categories of people based on characteristics.

ANITA ANAND: We can take another question?

KATHY NEW: Thank you. My name is Kathy New. My question is when this inevitably goes horribly wrong and the AI war crimes are up in Court, where will the responsibility lie? Will it be with the developers, the programmers, the designers or the people who are deploying these weapons and are they aware at the moment of what their responsibility may be in the future?

STUART RUSSELL: The question of responsibility is an interesting one. As I mentioned, there was a discussion of an accountability gap, and I don't think the gap is really there. If you deliberately target a civilian population then that's criminal intent. If you use a weapon whose outcome you can't predict reasonably accurately and it ends up killing lots of civilians, then that's criminal negligence, and so in both cases the party that programmes the mission into the weapon and launches it would be responsible.

There could be other cases where the mission that was programmed into it was in fact a legitimate mission but there was a software malfunction, and that could get a bit more complicated, and I'm not a lawyer so I'm not going to say exactly where that would come out. There would be then, I think, some shared

responsibility because you shouldn't be using weapons that haven't been properly tested.

ANITA ANAND: We have somebody who's working on the development of AI here, machine learning and defence and security, Dr Steven Meers is with us. This idea of the moral line and where to never cross it or go near it, how high is that in your mind when you're looking for answers to problems that exist for the army?

DR STEVEN MEERS: When we are trying to develop future concepts or future countermeasures to autonomous systems within defence, our kind of ethical and responsible approach really is at the forefront of what we do. We have ethicists that we work with who help us guide and develop our approach. In particular, we think very hard about the vulnerabilities and the kind of the misuses of the technologies that we develop, so we focus very hard on kind of responsible and ethical application that really saves lives and tries to reduce harm. So, absolutely it is front and centre of our approach.

ANITA ANAND: So, pushing a bit further, when you say "we," how many people are working on this? I mean, is there sort of like a hive mind? How many scientists are together pushing the frontiers on this?

DR STEVEN MEERS: Absolutely. So, for those of you that aren't familiar, DSTL is the Defence Science and Technology Laboratory, it's the science and technology arm of the Ministry of Defence. There are around four and a-half thousand scientists and engineers that are working at DSTL and we cover a very broad spectrum of technologies, from space systems to chemical and biological defence that you've mentioned, and also AI and data science, and within the AI and data science area we research a wide range of different technology areas.

A really important part of our research is about human machine teams. We really see the future of AI as being about augmenting the human capital that we do have using the machines to support the human decision makers and help them make sense of large quantities of data to do the things that the machines look at that, and to free up the human capital to focus on its strengths.

ANITA ANAND: Thank you very much. And there's a question I promised on that end?

RUBEN BOSS: Hi, my name is Ruben Boss and I'm curious as to whether you think there's any chance that these AI weapons that you are talking about could be used in espionage, or ways like that, to hide responsibility and be able to commit attacks where the country or organisation behind it is hidden and unknown?

STUART RUSSELL: I think that's a real concern and I know for a fact that many countries area worried about unattributable assassinations of their leaders, other politicians, which could be used to create internal conflicts in countries and all kinds of other mischief.

I think there's also potential uses by criminals as well. You could imagine a website where you upload the name, address and photograph of someone you want to get rid of and it's 49 pounds for one or 99 pounds for three, and this is not particularly desirable, and perhaps I'm exaggerating. But attribution is, again, something that nations can negotiate with each other. They can insist that weapons have markings of origin and that could help.

We've had similar discussions about attribution of nuclear explosions and attribution of Novichok, for example, in the UK and deniability is a real problem.

DAVID BALMAN: Hello. Professor David Balman here. I wonder if we assume that AI in general can develop the ability to mimic human emotion intelligence, whether we should be thinking about restricting or embracing that kind of capability in warfare AI in order to make the best decisions?

STUART RUSSELL: Yes. Many people cite human emotional responses as a problem, whether it's fear, or hatred, or revenge, that a lot of atrocities in war come from humans who are placed in these very difficult situations and their emotional response is inappropriate. And on the other side, as happened with Human Rights Watch, they cited human compassion as a check on the killing of civilians, and I think there's some validity to that, especially in the domestic situation. It's quite hard to get your soldiers to kill their fellow countrymen, fortunately, and many people worry that if autonomous weapons are available, they wouldn't have that same degree of resistance that they could be used, or

even just threatened to be used, as a way of controlling civilian populations, which I think could be very bad, so Amnesty International, for example, is quite concerned about that.

I don't know that AI systems are ever going to have real emotions but something resembling compassion, something saying, "There's something about this situation which doesn't feel right," and that this is not really an enemy, this is not really a threat and perhaps I should refrain, that might be a good thing.

ANITA ANAND: So far, we've spoken about a lot about the frontline and what happens actually in the heat of war. I'm kind of interested in what's happening behind that with policy and governance, and Dr Keith Dear is a Director of AI Innovation at Fujitsu Defence and National Security. I'm right in saying that you did previously work as an Expert Advisor on the Integrated Review, advising Number 10, among other places; is that right? Can you just tell me about the political will that you have come across? Name names, if you like, but I'll understand if you don't. Is there a great desire to push forward in using AI in lethal terms?

DR KEITH DEAR: No, I don't think anybody has a huge appetite to accelerate lethal autonomous weapon systems for the sake of accelerating lethal autonomous weapon systems. I think there are real concerns about international security and stability in the face of nations developing AI to support decisions at all levels, which with the vast volumes of data we have is going to be essential.

There are worries about how we best exploit those systems without damaging international stability, and there are concerns about other nations developing those things and your nation, not in the end, the international atmosphere is a competitive environment, and there are worries about what that means and how you might regulate the things that Stuart talked about. So, you might ban yourself from developing a system whilst another nation develops them and therefore, you're now vulnerable to the AI equivalent of nuclear blackmail. So, there are real concerns and I think it's important that we consider them.

ANITA ANAND: But Stuart's argument is a compelling one, isn't it, that after having a ban on chemical and biological weapons we don't have wars fought with those, do we?

DR KEITH DEAR: If you looked at the Cold War period, the Soviets had significant stores of chemical and biological weapons and NATO, fortunately, spent an awful lot of time training for how they would fight in a chemical and biological environment, because they fully expected that was what that war might look like. Now, we could have a longer debate about the nuclear peace and why it may have been that that war never happened, but it wasn't that there was nobody willing to deploy them. So, I'm not sure.

I think these are hugely complicated issues and we are right to be concerned and I think the debate that we're having is really important. I think it's also important that we consider just how competitive the international environment is and the risks of not having certain systems, and that we have the debate that we're having tonight.

ANITA ANAND: Thank you. Let's go and get a question over here?

JO HOOKER: Hello. Jo Hooker. Just about banning these weapons and pulling them back, isn't that a real challenge because, if I can put it this way, they're shiny, they're new, they're exciting, and so the chances of people drawing back from that are going to be very, very slim. When I say "people" I mean governments and other countries and so on and so forth.

STUART RUSSELL: I think there is that. The possible advantages, military advantages of these weapons are very clear to the military planners. For example, if you look at what happens in a dogfight between an autonomous fighter aircraft and a human-piloted aircraft, it's not very pretty.

Interestingly, when you look at how people thought of biological weapons, they really thought of it as possibly the future of warfare. They were designing, or at least trying to design, weapons that could wipe out only people of Slavic origin, for example, or they would provide antidotes to the weapon to their own population and then just wipe out everybody else. Those were the shiny new things, and the Russian government persisted, invested huge resources into growing their biological weapons programme.

So, those were shiny new things, but now I guess we know better, or at least we think we know better, and so opinions can change. We can decide that certain types of weapons, although they have military value, as long as there is a real agreement with real teeth and real verification and confidence that one side is not gaining a surreptitious advantage over the others, then we can have agreements that are actually beneficial to every country.

ANITA ANAND: Thank you. And the last question?

VIRGINIA WATSON: Hi. I'm Virginia Watson. I was wondering about the balance of AI and human – the power balance, what should it be, because AI obviously isn't perfect but nor are humans, and where does that fall?

STUART RUSSELL: I think I'm a human chauvinist in the sense that I think that human beings ought to have control for ever and, actually, the subject of the first and fourth lectures in the series is why that might not be the case and how we can try to ensure that it will be the case. And when I started working on this, I didn't know but I found out that there are actually people who would be completely happy for the human race to disappear. Some because they think that the human race has destroyed and pillaged the planet and they think that nature deserves to be protected against humans and we should just get rid of all humans. These are sometimes called the "anti-natalists." But there are other people who think that if machines are more intelligent than us then it's better that they control the earth and the future and not the human race.

But, actually, I don't think of this as a sort of IQ competition, whoever wins the IQ competition gets to rule the earth, because that's not what makes human existence valuable. Period.

ANITA ANAND: Well, that does sound like an ending to a programme to me. Thank you very much. Next time, as Stuart says, we're going to be in Edinburgh, and Stuart's going to be assessing how AI will change the way we work, how it's going to impact on jobs, what it will mean for the economy.

But for now, a big thanks to our audience, our hosts here at Manchester University, to our audience, and most of all, to our Reith Lecturer for 2021, Stuart Russell.

(AUDIENCE APPLAUSE)

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 3: AI and the Economy
Edinburgh

ANITA ANAND: Welcome to the third of this year's BBC Reith Lectures from the University of Edinburgh, with Stuart Russell. He's a Professor of Computer Science and founder of the Centre for Human Compatible Artificial Intelligence at the University of California, at Berkeley.

Now, in a series of four lectures Stuart has been examining what he says is the most profound change in human history, as the world becomes increasingly reliant on super-powerful AI.

Last time, Stuart discussed AI and Warfare. Now, he's going to be turning his attention to something which we will increasingly have to think about, the impact of AI on our jobs, on the wider economy, and if the robots really are coming, what's going to happen to us? I mean, it's one thing to have a machine make and drive cars, it is quite another to have a robot administer end-of-life palliative care, for example. And if the future is one without work, for most of us, what are we going to do? Lots to consider then.

Ladies and gentlemen, will you please welcome the 2021 BBC Reith Lecturer, Professor Stuart Russell.

(AUDIENCE APPLAUSE)

STUART RUSSELL: Picture the scene. It's a few days after Christmas 1964. A two-year-old boy is playing on the floor of his grandparents' house, while from one of his grandfather's many home-made vacuum tube radios we hear the bright predictions – a brisk and authoritative voice telling us that artificial intelligence, or cybernation, as the speaker calls it, would soon create a better world. I was that two-year-old, and Leon Bagrit was that voice, devoting the entire series of Reith Lectures to "The Age of Automation." Perhaps, unconsciously, the lectures influenced my career choice.

Now, Leon Bagrit was the head of Elliott, at that time the largest computer manufacturer outside the US, and he put forth many progressive and prescient ideas. He also complained that "For a long time we have been ruled by men who believe that only a good arts education matters," and he suggested that senior civil servants and ministers should be required to have at least some understanding of science. I think we're still waiting for that one.

Bagrit also predicted a massive reduction in employment, but he was by no means the first to do so. In a 1930 essay, "Economic Possibilities for our Grandchildren," the great British economist John Maynard Keynes introduced the phrase "technological unemployment" and predicted the end of employment within a hundred years. The Luddites, of course, were worrying about it in the early 19th century, and we can go all the way back to Aristotle in 350BC, who wrote:

"If every instrument could accomplish its own work, obeying or anticipating the will of others…if, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves."

So, the first question I'm going to discuss today is whether Aristotle and Keynes and Bagrit were right: will the progress of AI lead to the end of work?

The second question: is that a good thing?

Now, the first question, technology and employment, has been a staple of economic theory and debate. I am certainly not an economist, and any resemblance to real economics here is purely coincidental, but I could not ignore the topic. Parents often buttonhole me after lectures, wanting career advice for

their teenagers. The media are full of articles about AI and jobs. And for obvious reasons, governments all over the world are starting to pay attention. In the early twentieth century, technology put tens of millions of horses out of work. Their "new job" was to be petfood and glue. Human workers might not be so compliant.

I want to be clear upfront that when I'm talking about AI, I'm not just talking about AI as it exists today. Today's AI is certainly having an impact, and applications already in the pipeline, such as self-driving taxis, may have an even bigger impact.

But I'm talking about general-purpose AI, that is machines that can quickly learn to perform well across the full range of tasks that humans can perform. This, after all, has been the goal of AI since the beginning. We're not there yet, but if, as most experts believe, it's a plausible outcome in the next few decades, we must prepare for the potential consequences.

So, back to the first question: will the progress of AI lead to the end of work?

Aristotle says, obviously, yes! But economists know better - or at least, they thought they did. The classical theory for many years was that technological unemployment is impossible.

The theory says, and here I'm quoting a real economist, "Any advance that increases labour productivity also tends to raise the demand for labour, and thus employment and wages." You can make a mathematical model with too many Greek letters and prove this claim. There will be new jobs because the equations say so.

I think a double dose of motivated cognition, otherwise known as wishful thinking, went into this theorem.

First, there have been several technological revolutions in the past - agriculture itself, which replaced hunting and gathering, and then the mechanisation of agriculture and the mechanisation of industrial production. Yet still we have jobs! So, economists were motivated to come up with a model in which this counterintuitive outcome would necessarily be true.

Second, and maybe this is a cynical explanation, they had a strong desire to support technological progress against what they saw as the forces of obscurantism.

It became common to refer to the “Luddite fallacy” that machines were taking people’s jobs, and to the “lump-of-labour fallacy” that the amount of work to be done is fixed, so if machines do more, people do less. The word “fallacy” here ensures that no one can disagree.

Eventually, the economists admitted that the post-technology adjustments may take time, that some sectors may benefit more than others, and that the share of wealth going to capital rather than labour might increase - as in fact it has.

Still, there was a great reluctance to admit that any group could be harmed in absolute terms, despite the fact that low-skilled workers’ real earnings have declined substantially in many developed countries over the last 50 years.

To clear this up, let’s conduct a simple thought experiment. Let’s imagine that technology creates a twin of every person, and your twin shows up to your job - whether it’s your current job or one of those wonderful new jobs that will be created. Your twin is a bit more cheerful, a bit less hung over, and willing to work for nothing. How many of you would still have a job?

And you can see where the equations go wrong. Employment would be higher, it’s just that it wouldn’t be employment of humans.

Now, I think this debate has persisted so long because automation can increase or decrease employment, depending on circumstances, through both direct and indirect effects.

Consider, for example, what happens to housepainters as painting technology improves. I’ll summarize the technology level as the effective width of the paintbrush:

So, at the beginning, if the brush is one hair wide, a tenth of a millimetre, it takes thousands of person-years to paint a house and essentially no housepainters are employed.

With brushes a millimetre wide, perhaps a few delicate murals are painted in the royal palace by a handful of painters.

At one centimetre, the wealthy landowners follow suit and hundreds of housepainters are employed.

At ten centimetres or four inches, we reach the realm of practicality: most homeowners have their houses painted inside and out, although perhaps not all that frequently, and tens of thousands of housepainters find jobs.

Once we get to wide rollers and spray guns, the equivalent of a paintbrush about a metre wide, the cost to paint a house goes down considerably, but demand quickly saturates. No one needs their house repainted every few months, so the number of housepainters drops somewhat. But when one person manages a team of 100 house-painting robots, the productivity equivalent of a paintbrush a hundred meters wide, then whole houses can be painted in an hour and very few housepainters will be working.

Thus, the direct effects of technology work both ways: at first, technology can increase employment by reducing costs and increasing demand; subsequently, further increases in technology mean that fewer and fewer humans are required once demand saturates.

The economist James Bessen has called this the inverted-U curve: as technology progresses in a given sector, first employment goes up, and then it goes down. Bessen catalogues several major industries showing exactly this pattern.

In other words, the direct impact of any particular technological advance depends on where you are on the curve. For example, if we all lived in very big houses, switching from paintbrushes to rollers would increase the numbers of painters, because there'd be lots of unmet painting need that could now be met at the new low price. But if we all lived in very small houses, the need for painting would already be met by paint brushes, and rollers would decrease the number of painters instead.

So much for the direct effects. What about the indirect effects described by the techno-optimists?

Well, obviously, some people have to make the painting robots. How many? Far fewer than the number of housepainters the robots replace - otherwise, it would cost more to paint houses with robots and not less, and no one would buy the robots.

And then, and this is something that's often missed in the debates, because we pay less for house painting, we have more money to spend on other things, thereby increasing demand and employment in other sectors. I'll call this "the wealth effect."

Economists have tried to measure the relative sizes of all these effects, but the results are inconclusive. Instead, they have tended to fall back on the “big picture” view: automation increases productivity, so, as a whole, we enjoy more goods and services for less work.

Back in the real world, however, we also see what the economists Erik Brynjolfsson and Andy MacAfee call the “Great Decoupling”: a doubling in productivity since 1970 but only a tiny increase in median real income, even though these used to move in lockstep. And this is associated with the Great Hollowing Out - many middle-class jobs disappearing and being replaced with low-wage, low-security jobs. And of course, a substantial shift in income share from labour to capital and to the topmost echelons.

Looking forward, it seems likely that the technologies under development will continue this trend. The most obvious is the self-driving taxi or goods vehicle. This has taken longer than some people expected, but it’s now starting to happen.

In the coming decade I think we’ll see real advances in language understanding. Machines will be able to interpret the content of human communication sufficiently well to automate many short-interaction tasks - customer service, insurance claims, and so on. The low-level programming jobs and computer-based clerical tasks typical of outsourced work are also likely to disappear as the technology of robot process automation advances.

But for a long time, the techno-optimists pointed to complex sensorimotor tasks such as folding towels as beyond the scope of AI. Even Andy MacAfee, who has been a leading thinker on the somewhat pessimistic side, held onto this hope, until I showed him my colleague Pieter Abbeel’s video of a robot neatly folding a pile of laundry. “Oh bleep,” he said, “That’s another 500 million jobs gone.”

Some of the optimists say that we can just develop technologies that augment humans rather than replacing humans, or, as economists describe it, complement rather than substitute. They love to use healthcare as an example, because, well, who could argue with better healthcare? An automated radiologist, for example, could take care of the routine cases, while the human expert focuses on the interesting, difficult cases and has time for some soft-focus, sunlit empathy with the patient in the PR video.

But, as we saw in the case of switching from paintbrushes to rollers, technologies that improve productivity are not intrinsically complementing or substituting; it all depends on the demand response. I’m pretty sure that if

radiologists become more productive using AI, we will not break our legs in more interesting and difficult ways just to keep the radiologists in work.

So, we cannot just tell the AI researchers to develop “complementing technologies.” We have to identify areas where raising productivity or quality and lowering costs would lead to much higher demand. This really means finding unmet needs - tasks that are just too difficult or expensive or dangerous for people to do well - and meeting those needs.

It’s not too hard to think of possibilities, such as removing graffiti, cleaning up the environment, inspecting shipping containers, fighting forest fires in California, and improving the quality of education through personalised instruction.

Certainly, pursuing applications of AI that meet unmet needs will put us in a better position for the long-term future.

But in that future, general-purpose AI will push essentially all sectors into decreased employment - onto the downslope of the inverted-U curve. Put another way, with general-purpose AI, the house painting example is a metaphor for the entire economy, not just one sector. For almost every task, there will be the equivalent of a team of house-painting robots. Even though global wealth will increase enormously, it’s not obvious that there are “other sectors” where the wealth effect can lead to compensating employment gain.

Now, this isn’t necessarily a bad thing. For the last 10,000 years or so, our societies have used most people as robots, performing repetitive physical and mental tasks, so it is perhaps not surprising that real robots will soon take on those roles. If you were a science fiction writer in about 8000 BC and you wrote that one day, people would get up, go into windowless buildings, do the same thing 10,000 times a day, and do this nearly every day until they died, your nomadic hunter-gatherer colleagues would think you were nuts. That couldn’t possibly happen. But it happened. Now, it’s coming to an end. How do we respond?

It seems to me inevitable that we will need fundamentally new socioeconomic arrangements. But economists mostly tweak the knobs on the current system: raise this tax a bit, subsidise retraining, and so on. They’re just not in the “invent a new economic system” business.

Science fiction writers, on the other hand, are in that business, but they tend to dream up arrangements that would immediately fall apart, according to the economists, because the economic incentives wouldn’t cohere. Either that or

their stories rely on warp drives to provide for endless expansion across the universe.

So, my evil plan, hatched with the World Economic Forum in 2019, was to lock the economists and science fiction writers into a room and not let them out until they came up with a workable vision of an economy where AI does most of what we currently call work.

Unfortunately, they were saved in the nick of time by Covid.

We had several Zoom workshops instead. That didn't solve the problem, but they did reveal an interesting split in how people answered my second question for this evening -would the end of work be a good thing?

One camp largely agreed with Keynes who said, in his 1930 essay:

"Thus, for the first time since his creation man will be faced with his real, his permanent problem - how to use his freedom from pressing economic cares, how to occupy the leisure, which science…will have won for him, to live wisely and agreeably and well."

Now, modern proponents of Keynes's vision usually support universal basic income, or UBI. UBI provides a reasonable income, derived from tax revenues, to every adult, regardless of circumstance, allowing people to spend their time as they see fit.

Keynes made a clear distinction between those who strive and those who enjoy - those "purposive" people for whom "jam is not jam unless it is a case of jam tomorrow and never jam today" and those "delightful" people who are "capable of taking direct enjoyment in things."

He goes on to say that "striving" is one of the "habits and instincts of the ordinary man, bred into him for countless generations" rather than one of the "real values of life." He predicts that this striving instinct will gradually disappear. But he does hint at the need to prepare people for their life of leisure:

"It will be those peoples, who can keep alive, and cultivate into a fuller perfection, the art of life itself…who will be able to enjoy the abundance when it comes."

Perhaps he's thinking of the French and Italians, in contrast to the British.

Leon Bagrit also thought that the end of work would be a good thing, but he went much further than Keynes in his insistence on the need to "develop people capable of living the fullest possible lives in an age of plenty," thinking that it wouldn't happen automatically. He devoted an entire lecture to the question of educating people for this new world.

So much for the view that the end of work is a good thing.

The second camp in our Zoom workshop believed the opposite: work - in the sense of an organized system of mutually beneficial exchange - might still be essential. Universal basic income represents merely an admission of failure. It assumes that most people will have nothing of economic value to contribute to society. They can be fed, housed, and entertained -mostly by machine - but otherwise left to their own devices.

This second camp also tended to believe, contrary to Keynes, that striving will not and should not disappear; indeed, it is intrinsic to what it means to be truly human. Striving and enjoying are not mutually exclusive, they are inseparable: true enjoyment and lasting fulfilment come from having a purpose and achieving it, or at least trying, usually in the face of obstacles. There's a difference between climbing Mount Everest and being deposited on top by helicopter.

So, what value could people contribute in an economic sense in the Age of Automation?

The inevitable answer seems to be that people will be engaged in supplying interpersonal services that can be provided - or which we prefer to be provided - only by humans. That is, if we can no longer supply routine physical labour and routine mental labour, we can still supply our humanity. We will need to become good at being human.

Now, this is one area where we have a comparative advantage over machines. For example, I don't need a PhD in neuroscience to know what it feels like when you hit your thumb with a hammer. I can just hit my thumb with a hammer to find out. And most of us already know what unrequited love feels like - so we don't need to try it again - and we can sympathize appropriately.

Current interpersonal professions include psychotherapists, executive coaches, tutors, counsellors, social workers, companions, and those who care for children and the elderly. The phrase "caring professions" is often used, but that's a bit misleading: it has a positive connotation for those who provide the care, but a negative connotation of dependency and helplessness for the recipients.

But we're really talking about "perfecting the art of life itself." This is not dependency but growth. The capacity to inspire others and to confer the ability to appreciate and to create - be it in art, music, literature, conversation, gardening, baking, or video games - is likely to be more needed than ever.

Now, if we accept the proposition then that most people will be engaged in the interpersonal professions, this also raises the question of income distribution. Childcare, for example, is poorly paid and poorly regarded, partly because we don't really know how to do it well.

Contrast this with, say, orthopaedic surgery. I hope we wouldn't just hire bored teenagers to repair broken bones at five pounds an hour plus all they can eat from the fridge, as we do for babysitters. Orthopaedic surgeons are highly paid and highly respected because centuries of research underlie their training, and it actually works. It is a high-value-added profession. Childcare is not always high-value-added; in some cases, including the babysitter who very kindly tried to teach my seven-year-old sister and my nine-year-old self to smoke, sometimes it's value-subtracted.

Unfortunately, our scientific understanding of the individual mind and its growth and fulfilment is weak at best. We simply don't know how to add value to each other's lives in consistent, predictable ways. Partly because individuals are all so different. We have had moderate success with some psychiatric disorders, but we are still fighting a Hundred Years' Literacy War over something as basic as teaching children to read.

This suggests a need to retarget our education system and our scientific enterprise to focus not on the physical world but on the human world. It sounds odd to say that happiness should eventually be an engineering discipline, but that seems to be the inevitable conclusion.

Such a discipline would build on basic science - tools that provide a better understanding of how individual human minds work at the cognitive and emotional level - and it would train a wide variety of practitioners, ranging from life architects, who help individuals plan the overall shape of their life trajectories, to professional experts in topics such as curiosity enhancement and personal resilience, to professional lunchers. If based on real science and effective training, these professions need be no more "woo-woo" than bridge designers or orthopaedic surgeons are today.

So, it seems that whether we think of the end of work as a good thing or a bad thing, either way we need a radical redirection of our science and education:

either to equip individuals to “live wisely and agreeably and well” or to support a human economy based largely on high-value-added interpersonal services.

Reworking our educational and research institutions in this way will take decades - I am reminded of the fact that it took Oxford University 125 years to approve a proposed degree programme in geography, by which time it may have been too late - so it’s a good idea to start now and a pity we didn’t start long ago, as Leon Bagrit suggested in 1964. The final result, if it works, would be a world well worth living in.

Without such a rethinking, we risk an unsustainable level of socioeconomic dislocation. Faced with this prospect, some leading economists, far from pronouncing technological unemployment impossible, are now thinking of joining the Luddites in advocating for a slow-down in the development of AI, so that we can have a chance to work this out. I’m not sure that slowing down AI is possible, but I can certainly anticipate a time when we reserve interpersonal professions and tasks for humans. If it’s more blessed to give than to receive, we must not, by receiving everything from machines, cut off humans from the opportunity to give.

Thank you.

(AUDIENCE APPLAUSE)

ANITA ANAND: Stuart, thank you so much. It’s absolutely fascinating to me that you are an expert in artificial intelligence, and you have been taking us on this relay race from Bagrit to yourself, to who knows what, and all I kept thinking about is what does this say about the human condition, what it is to be human, and you laid out very clearly what people think might be the eventuality of having the end of work. I want to know what you think it will be like.

There are two scenarios. One, I lose every excuse not to learn the piano, and I’m much nicer to my children, and I catch up on all of the wonderful things on Radio 4, or that terrifying Dystopia, the animation WALL-E where we all sit on our bottoms getting fatter watching infomercials. I mean, what do you think the human condition leans towards?

STUART RUSSELL: This is a great question. I think it’s really the fundamental question and up until now human race hasn’t really had a choice. We have to get up and get out of bed otherwise we starve to death. And one economist actually described economics not as the study of money and so on but as the study of why people get up and get out of bed. I thought that was a great description of what economics really is. But if that incentive is taken away, one

might imagine, as I think happens in WALL-E, that we lose the incentive to spend 20 years in education.

We have to do that now because otherwise our civilisation would fall apart, right? If the next generation can't run the civilisation, then it falls apart. So, we spend actually, if you add it up, about a trillion person years teaching the next generation how our civilisation works, everything we know and maybe some more, and if that incentive goes away, if we can now hand over the running of our civilisation to machines, why would anyone bother?

There's a television series called Humans where the daughter asks the dad, the dad wants her to go to medical school, the daughter says, "Why would I bother? It takes me seven years to get through medical school and the robot can learn it all in seven seconds," and so it's not obvious that with UBI and with unlimited wealth that we would still learn everything that the human race has learned so far and push the boundaries of knowledge forward and develop internally a strong sense of purpose, and mission, and go out there and interact with lots of other people.

I think some people might, but the evidence is pretty mixed so far, for example, how aristocrats in the 18th and 19th centuries spent their time. There was a lot of dissolute behaviour, and I think that's one of the reasons why the brutal British boarding school was invented.

ANITA ANAND: Well, I mean, it's a very good time to interact with other human beings now while we can. Let's open this up to questions from the floor. Sir, over to you?

PAT KANE: Hi. Thanks very much, Stuart. My name is Pat Kay and I'm a writer and musician. I'm slightly horrified by your vision of a sort of science of conviviality and happiness. I would have wanted this technology to make humans more free, and so your issue is are we capable of that freedom. But do you think about the fact that at the same time as AI is threatening to upend our economies, people from the IPPC and the climate crisis community are saying we're going to have to produce less and consume less? So, we seem to be in this either terrifying moment of division and polarity or this wonderful moment when human history actually begins. We have the resource to be as free and agentic, as evolution clearly designed us to be.

STUART RUSSELL: So, many people believe that AI will be the thing that solves the climate problem. I don't believe this at all. I think the climate problem is up to us. It's what economists call a coordination problem, and no-one wants to be the first to pay all the sacrifices when the other countries are busy pumping

out the carbon dioxide. The vision that an AI-powered economy could be much more productive relies on certain assumptions about the ability of AI to make the economy simultaneously sustainable. For example, to go from 20 per cent recycling, or whatever we have now, to something more like 98 per cent recycling. I think that is a feasible improvement that we could make because recycling is incredibly labour intensive and not very remunerative and so we don't do most of it.

But the things that are finite - land, for example – that's a real problem. I was looking at Greenwich, Connecticut, which is a completely urbanised area in the United States but a very wealthy one. So, if you said, "Let's have a Greenwich, Connecticut standard of living for everybody," if you just had as much land per person as they have in Greenwich, Connecticut, the Chinese population would cover every square metre of the earth. Ignoring all the other countries, just China, that would cover the whole earth. So, land is going to be, for the foreseeable future, the bottleneck resource for any general increase in the standard of living and we have to, I think, learn to live on top of each other happily.

ANITA ANAND: You said you're a musician. Can I just ask you whether you think AI will enhance human creativity or replace it or crush it? I mean, what is the relationship going to be, do you think, between creators, musicians, writers and AI?

PAY KANE: Well, I read the other week that Beethoven's Unfinished Symphony was partially finished by an AI, along with some musicologists, and the experts couldn't tell the difference between the bits that were AI and the bits that were human. I'm actually genuinely interested in the question about whether if AI starts to generate itself, which I'm sure you know these evolutionary models, when will it not just mimic own intelligence but could it possibly come up with a completely different intelligence, because that's what artists are always looking for, difference and a uniqueness. May AI begin to communicate to us a kind of intelligence that we've never experienced before? Now, as an artist, I'm excited by that.

ANITA ANAND: You're excited. Okay. I can see in the audience we've got Val McDermott, who is a great writer. Are you as excited? Is this going to set you free or are you worried about this, Val?

VAL McDERMID: I think it's got lots of fascinating potential, but I'm really interested in the fact that aspiring writers are always told to write what you know, which I've always taken to mean what you know psychologically, emotionally and not just what you know practically in terms of the job you've done or the places you've lived. Will AI have the kind of experience that will make

fiction possible, or will it simply be a regurgitated sort of algorithmic amalgam of what's already out there?

STUART RUSSELL: That's a really good example because I think that, as I said, we can hit our thumbs with hammers and have the first-person subjective experience because our nervous systems are very similar, right, we're fairly sure that what it feels like to you is what it feels like to me, and machines absolutely have none of that. They don't have, as far as we know, any subjective experience, and so they'll never be able to talk about what it felt like to have unrequited love. All they could do, at best, was try to guess how to say what it feels like for humans and so it wouldn't be the same writing as it is for humans, and this is one area where, as I say, I think we have a comparative advantage and we're likely to keep it.

ANITA ANAND: Well, you say that, but this is a section of a completely AI-written poem. Let me just read a bit. It started off – it was last year, I think you've probably come across this, completely generated by AI, starts off with gibberish and then you have this section in the middle:

"We travel across an empty field in my heart. There is nothing in the dark, I think, but he. I close my eyes and try to remember what I was. He says, 'It was an important and interesting day because I put in his hand one night a box of light that had been a tree.'"

Now, I put it to you, ladies and gentlemen, that this might have been a very pretty poem had you thought it was written by a human being. This AI hasn't fallen in love, has not lain in a field, has not had unrequited love, and yet that sounds pretty poignant. Does it matter that it hasn't?

STUART RUSSELL: We can still consume it. I mean, sometimes on the beach the waves throw up pebbles in beautiful patterns, right, but that doesn't make it art. It's still a nice thing to look at but it isn't art in the same sense, and I think this poetry is more like what the waves are doing with pebbles on the beach rather than coming from any internal experience. It's a random assembly and regeneration of patterns of words that humans have produced in the past.

ANITA ANAND: Thank you. Let's take the question over there. There was a woman with a red mask?

HELEN BLACKBURN: Hi. My name's Helen Blackburn. I've been a mental health nurse for 33 years now and in that time I've worked with a lot of patients. You made reference to healthcare, and I just wondered, do you really think that a robot or a machine will ever be able to read things like body language, eye

contact, tone of voice and to understand empathy and to be able to deliver high-quality, person-centred care?

STUART RUSSELL: I think it's difficult. There is some evidence that in some kinds of psychological conditions machines that have discussions with you are actually preferred by patients. They find it easier to talk to the machine, the machine is not embarrassing them or threatening to them in any way. But I think those are fairly limited circumstances. And again, the human doctor or psychiatrist or psychologist, because of the way our nervous system works, whenever we experience the presence of another person, as you say, their speech, their fidgeting and so on, we automatically, in some sense, create internal copy of that and our brain comes up with sort of an emotional underpinning of the explanation for why is that happening.

It's as if we perceive the emotions directly, even though we are just perceiving sound and body movement and so on, but we feel as if we're perceiving the emotions directly of sadness or uncertainty or shyness or whatever it might be, and I don't think the machines can have that. They can learn by those superficial statistical regularities from thousands, and thousands, and thousands of patients, when there's a quivery tone and the patient's not looking the doctor in the eye, et cetera, et cetera, et cetera, then it's probably this, but these are just superficial associations, there's no meaning to this from the machine's point of view.

ANITA ANAND: Let's take a question from over there.

DAVE WALKER: Hello. I'm Dave Walker. Husband, father, teacher. Not all in the same order. Perhaps a problem or a solution; do you ever see a point in the future where everything is turned on its head and in fact, perhaps from a teacher's point of view, it's very difficult to know that you are training or educating a person for the right job and to be able to get the best out of someone? Is there some point in the future where AI learns how to identify the humans who would be best for the jobs of the future?

ANITA ANAND: Sort of a mechanised Sorting Hat.

STUART RUSSELL: Well, I'm afraid to say that's already happening, to a large extent. So, AI systems are sifting through tens of thousands of CVs that are submitted for job applications and picking out the ones who are going to be interviewed, so it's already there.

The whole intersection of AI and education is a really interesting area. I think it's very much underexplored and we know, for example, that if you tutor a

child individually and you have a skilled human tutor, they can learn about three times as much as they do in a normal classroom. So, just think of that incredible waste of talent and potential that's happening, not just here and there but everywhere in the world. The only way you can have an individual tutor for every child, right, just the numbers don't work out if it's going to be a human so it has to be an AI system, and that's something that we could take on, as a field, as a grand challenge, and I think it would be of enormous value to the world.

I still think we would need humans to work with the children who are being tutored, addressing their emotional needs, social needs, doing more high-level guidance and planning for their education, but if that worked, if we could actually deliver an education of extremely high quality, I think we would be willing to hire more human teachers as well because we would be so happy with the potential outcomes that we could get.

ANITA ANAND: I mean, it's interesting, I sometimes do this but when I recognise somebody in the audience who might be able to shed some light about what goes on in governments, maybe not now but in the past, we have Douglas Alexandra, I see you in the audience, a former Labour Cabinet Minister in Gordon Brown's government. How much does this idea – I mean, Stuart's saying, "When they come to us, we'd better be ready," are you aware of politicians being even remotely close to even asking the question and recognising that AI is going to have a profound effect on our economies?

DOUGLAS ALEXANDER: I don't think it has yet moved beyond an understanding that it's something people need to understand. It's on the horizon but it's at the back of politician's mind at the moment, not at the front of politician's mind, in my judgment.

Stuart, thank you for the lecture. You talked about AI in education but let me ask whether there are ways that artificial intelligence can actually help policy makers reduce inequality or is it inevitably simply going to deepen existing divides within our economy?

STUART RUSSELL: So, I think education is one area where we can reduce inequality because if you can develop high-quality tutoring systems, for example, then they can deliver the best possible education at next to no money, and I think the places that need it most would be developing countries where most kids only get a year or two of school, if that, and their parents often can't afford it at all.

I do take your point that the status quo, if it continues, is going to exacerbate inequality because we just see a higher and higher share of wealth going to capital and leaving labour, and that's something that every politician

should be concerned about. I don't know any near-term solution other than redistribution.

ANITA ANAND: On this political point, because it fascinates me, what do politicians do about it, and one of the things that you were talking about was this universal basic income and what that would do to equality and inequality. Can we hear from an economist? I know you wanted to lock them in a room pre-Covid. We have got an economist here. Yes?

STUART RUSSELL: Are you volunteering to be locked in the room?

ANITA ANAND: Well, we may not need to go that far. But, I mean, as far as you're concerned, this idea of a UBI, a universal basic income, sets us free or not? Tell us who you are and who you represent?

EMMA CONGREVE: Hi. I'm Emma Congreve, I'm an economist at the Fraser of Allander Institute at the university of Strathclyde. Thank you very much for your lecture.

So, universal basic income, yes, it's got its fans, definitely, but it also has its issues. The big issue that would have to be faced is the cost of it and our institute did some research and for Scotland, came up with a figure for a sort of minimum adequate standard of living of around 50 billion pounds a year, that's about three times what we spend currently on social security here, and in order to fund that we looked at what you'd have to do to income tax, and we were talking about over 50 per cent starter rates going up to 85 per cent marginal rates for high income tax payers, and so there would have to be a transformation in people's understanding and willingness to pay for people not to work, I suppose, and that would take a very different mindset and would people be willing to still go out to work and pay that?

And the other issue, which is really important, I think, is around whether there will still be choice in this decision, whether we will create some kind of underclass that's just shut out of the labour markets. So, yes, how will we change people's mindsets, not just in terms of not working but actually those who do work having to fund those who don't.

STUART RUSSELL: There's also the problem of if people then just drop out of the workforce and stop paying taxes then you get a sort of vicious circle, and then you have to raise the tax rates even higher on those who are still working, and that doesn't work. And I think it also comes back to your view of human nature and experiments that have been done, there are several. They seem to have, I think, fairly positive results; people are not just playing

videogames, they're learning new trades, woodworking and baking and sausage making and so on, they're engaging in constructive social activities. But those are people who grew up in a work where purposive activity and goal-directed education is the norm and they have been through our schooling system and most of them had already worked. That's not a good experiment because what we're asking is, would it work if people stopped bothering with school because they don't intend to get a job when they're older, and if they had never worked would they have the mental discipline and sort of goal structures and so on to actually organise their lives in constructive ways, and those experiments simply haven't been done.

I have to say I'm a little bit in the second camp that I describe where I don't think, at least for the foreseeable future, that we could have a successful society where almost nobody was working.

ANITA ANAND: Thank you very much. Question over there?

KENZA: Hi. I'm Kenza. I'm an AI undergrad at the University of Edinburgh, and my question is more related to what we call the singularity or a hypothetical point in time where AI grows so much and grows so fast that we lose control over it. My question is-----

ANITA ANAND: You're talking – just for those who don't do what you do, we're talking about every sci-fi, horror movie ever?

KENZA: Yes.

ANITA ANAND: Right. Okay, just so we're on the same page.

KENZA: So, my question is, do you think that this is actually possible, there is a possibility that AI creates some kind of intelligence that we, ourselves, are not capable to understand or do you think that because it lacks those five tiers of subjectivity that it is not a possibility?

ANITA ANAND: Yes, how many minutes to Blade Runner?

STUART RUSSELL: Actually, I Am Mother, if you haven't seen that film, that's even more scary. So, this is actually the subject of the fourth lecture, so I don't want to give away too much of the story, but yes, absolutely it's possible. I think my friend, Max Tegmark, who's a physicist, has a good way of putting it, which is, "Why on earth would we think that no physical arrangement of atoms in the universe could possibly outperform the human brain?" If you thought that,

and this is one of the pushbacks that we get is, “Oh, what if humans are the most intelligent things in the universe?” Well, come off it.

Absolutely we should avoid methods of creating AI that result in things we don’t understand. We have to have rigorous mathematical understanding of what we’re doing and the more powerful it gets, the more rigorous that has to be.

MARY GALBRAITH: Hi, there. My name’s Mary Galbraith. I work in global technology transformation programmes. What I’d like to find out from you is what you think there is in this for females because there’s a big inequality already, we’ve spoken a bit about that, and at the moment women are mainly doing the caring professions, interpersonal care, that you’re telling us is going to be the future of work. So, will it be women who are the main earners, the main workers? What does it also do for men? And, equally, there’s another issue around AI and the algorithms that we know about that have already been discovered to have biases against females?

STUART RUSSELL: So, let me deal with the second question first. You’re absolutely right, there have been several documented cases where fairly extreme biases. So, Amazon, with their résumé filtering or their CV filtering algorithms, was actually just filtering out any résumé that had the word “woman” on it, and it was because the algorithm had been trained with datasets that were biased because of history. Particularly, they mostly hired computer scientists in the early days, they didn’t have a big marketing department or any of those areas that are more evenly balanced, and so the algorithms just reflected or even exacerbated the bias in the datasets. We now have a fairly good understanding of how that happens and how to prevent it from happening.

On the first question, I think you’re right. The caring professions that I described have been more attractive to women in the past. And I want to just apologise, I’m sure many of you want to go and look at Leon Bagrit’s Reith Lectures, which are pretty amazing and very prescient, but just beware that he has fairly old-fashioned attitudes about men versus women.

So, I don’t see any particular inherent reason, I think things are changing dramatically. I don’t think that women are particularly more suited for one kind of profession and men are more suited for another, I think it’s almost entirely, or entirely, the result of how we are socialised growing up. And where I teach at Berkeley, we’ve seen a real sea change just in the last decade or so in who’s going into PhD programmes. I think probably seven of the last 10 professors we’ve had have been female. My best ever student, Daphne Koller, was streets ahead of any of the male students I’ve ever had, so, I just think we have to be a little bit patient

but push hard to get rid of the attitudes that have caused the problem in the first place.

ANITA ANAND: And the final question?

JAMES KILLEEN: James Killeen. I'm a civil engineer, transportation planner. My question's completely unrelated to my field, but in terms of the impact on the economy. So far, we've focused on the developed world, and I wonder whether the developing world will be able to gain advantages at the same rate or align with us or whether we're just a risk of increasing the gap or maintaining the gaps?

STUART RUSSELL: It's a great question and, once again, I'm not an economist but what I understand from my economic colleagues, particularly those who work in development like Jeffrey Sachs, for example, is that there's a serious worry that advanced robotics will actually cut off the root that countries have used to develop their economies, which is the export of manufactured goods with relatively low-cost labour. If you think about China, for the last 20 years, Vietnam, Thailand and so on, their growth has come through that root and we may cut that root off, almost inadvertently, by re-onshoring manufacturing with advanced robotics instead of human workers, and so we have to think about what's the alternative development path for these countries and how could AI be helpful in that direction.

ANITA ANAND: And we're running out of time. We have almost run out of time, but I am desperate to ask you one final question. How long before we have an AI Reith Lecturer?

STUART RUSSELL: There's several of these milestone targets that people have come up with. One that's actually been very popular, it's called RoboCup Soccer, and the goal of that whole competition, when it started, was that I think by 2050 we'll have a team of robots that can beat the world champion soccer team, the world champion human soccer team. I think they have to make them soft, otherwise it won't be very nice to tackle.

And then the same group actually came up with another idea, which is the Nobel challenge, right, that we're going to try to get an AI system to win a Nobel Prize. And interestingly, so DeepMind recently developed a method for predicting how a protein will fold. So, given the amino acid sequence, predict the actual structure of the protein. That's an amazing breakthrough and they won this competition and then they simply published predictions for the folding of every single protein that the human genome produces as a public good. So, I thought that was really great, and perhaps they'll win a Nobel Prize for that, I

don’t know, but clearly the robot Reith Lecturer is much more difficult than winning a Nobel Prize.

ANITA ANAND: Oh, of course it is.

STUART RUSSELL: So, it’s going to be several years after that.

ANITA ANAND: Yes, but you have made it look easy. Thank you so very much. That is all we have time for. Next time, in his fourth and final lecture, Stuart will be pulling everything together and is going to give us all the solutions to all of the problems. Very simple. Looking forward to that, but for now, thank you to Edinburgh University, to our audience and to our Reith Lecturer, Stuart Russell.

(AUDIENCE APPLAUSE)

Lecture 4: Beneficial AI and a Future for Humans

Newcastle

ANITA ANAND: Welcome to the fourth and final BBC Reith Lecture of 2021 with Professor Stuart Russell.

We're in Newcastle, at the National Innovation Centre for Data, set up two years ago with funding from the government and Newcastle University. It's based in this state-of-the-art Helix science district on the site of a former coalmine. I mean, you could say, from coalmining to datamining if you like. It is symbolic of the changes the north-east of England have undergone.

The NICD's mission is to transfer data skills to the UK workforce. Current projects include using AI to help improve patients' walking and track endangered species. It is an ideal place to wrap up this year's series called "Living with Artificial Intelligence."

So far in his lectures, Stuart has outlined some of the major challenges artificial intelligence poses to our lives; about the way we work, how we wage war, and now, in this final lecture, Stuart offers us some solutions, some ideas how about how we might live with AI.

So, let's hear them now, will you please welcome the 2021 BBC Reith Lecturer, Professor Stuart Russell.

(AUDIENCE APPLAUSE)

STUART RUSSELL: Thank you, Anita, and thank you to the BBC for inviting me. It has been a delight to give these lectures.

Now, for those of you who are following the series, you may remember that I left you at the end of the first lecture with a bit of a cliff-hanger. The scene I described was one in which all human beings are passengers on a bus that is speeding towards the edge of a cliff.

That cliff is the loss of control over increasingly intelligent machines, as predicted by Alan Turing in 1951, when 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."

The speed of the bus comes partly from the potentially enormous benefits of general-purpose AI, that is, machines that can quickly learn to perform well across the full range of tasks that humans can perform. In the first lecture, I gave a very rough, low-ball estimate of the cash value of general-purpose AI at ten quadrillion pounds. That prize creates a lot of momentum.

I also mentioned a few of the reasons the various "sceptics" have given for paying no attention. One I didn't mention is perhaps the worst excuse of all: some AI researchers - after 70 years of insisting to the naysayers that AI is possible - are now saying there's no need to worry because we won't actually achieve general-purpose AI.

This is like the bus driver speeding towards the cliff edge saying, "Don't worry, we'll run out of petrol before we get there." This is no way to manage the affairs of the human race.

Now, it has been pointed out, correctly I think, that there's too much doominess these days - in climate, in politics, and particularly in predictions about AI. A couple of years ago I received a phone call from a film director who wanted me to be an expert consultant for a new film about super-intelligent AI. He, too, complained about doominess, so my job would be to explain how the human protagonists in the film could outwit the super-intelligent AI and save humanity. "Sorry, they can't," I said. And so ended my career in films.

My task today is to dispel some of the doominess by explaining how to retain power, forever, over entities more powerful than ourselves - entities that we cannot outwit. I'll call this the control problem.

To solve this problem, we'll have to go back to the very beginning, the core of how AI is defined. Machines are intelligent to the extent that their actions can be expected to achieve their objectives. Almost all AI systems are designed according to this definition, which requires that we specify a fixed objective for the machine to achieve or "optimise".

The problem with this approach was pointed out by Norbert Wiener, the founder of cybernetics, in 1960. He said:

"If we use, to achieve our purposes, a mechanical agency with whose operation we cannot interfere effectively we had better be quite sure that the purpose put into the machine is the purpose which we really desire."

And there's the difficulty: if we put the wrong objective into a super-intelligent machine, we create a conflict that we are bound to lose. The machine stops at nothing to achieve the specified objective.

Suppose, for example, that COP36 asks for help in deacidifying the oceans; they know the pitfalls of specifying objectives incorrectly, so they insist that all the by-products must be non-toxic, and no fish can be harmed. The AI system comes up with a new self-multiplying catalyst that will do the trick with a very rapid chemical reaction. Great! But the reaction uses up a quarter of all the oxygen in the atmosphere and we all die slowly and painfully. From the AI system's point of view, eliminating humans is a feature, not a bug, because it ensures that the oceans stay in their now-pristine state.

So, there's little chance that we can completely and correctly specify the full objective, the one that matters - that is, humanity's ranking of all possible futures. We need a different way of thinking.

Now, early in 2013, I was on sabbatical in Paris, and I spent a good part of that time thinking about this problem. I also joined the chorus of an orchestra, L'Orchestre Lamoureux, as a very amateur tenor, and one evening I was on the Métro heading to rehearsal and listening on my headphones to the piece I was learning, Samuel Barber's Agnus Dei. "This is so sublime," I was thinking to myself, and, as one sometimes does in Paris, thinking "Live for this moment," even if the rest of the time in Paris one is thinking, "This moment is frustrating and humiliating."

But then, as often happens, my day job spoiled the moment, and I wondered how on Earth an AI system could ever know what constituted such moments - whether sublime or frustrating or humiliating - for a human being.

And then it occurred to me. We have to build AI systems that know they don't know the true objective, even though it's what they must pursue.

And all the other consequences came rushing in, including the fact that this would solve the control problem.

Over the next few days, partly in deference to the Three Laws of Robotics proposed by the great science fiction writer Isaac Asimov, I wrote these ideas down in the form of three principles.

The first two essentially say what I just said. The first principle is that:

- The machine's only objective is to maximise the realisation of human preferences.

So, the machine will be purely altruistic towards humans, with no objectives of its own, including self-preservation as commanded by Asimov's Third Law.

I need to clarify the word "preferences" here. These are not simple preferences - what kind of pizza you like. Nor are they expressed preferences - you want to be ruler of the universe and so on.

Instead, imagine watching two films, each describing in sufficient detail and breadth a future life you might lead, including everything that you might care about; and deciding which of these two futures you prefer. Now, technically it's a bit more complicated than that, but that's the basic idea.

The second principle is that:

- The machine is initially uncertain about what those preferences are.

This is the core of the new approach: we remove the false assumption that the machine is pursuing a fixed objective that is perfectly known. This principle is what gives us control, for reasons that will become clear soon.

The third principle is that:

- The ultimate source of information about human preferences is human behaviour.

Here "behaviour" means everything we do, which includes everything we say, as well as everything we don't do, such as not reading your email during

lectures. It also includes the entire written record because most of what we write is about humans doing things - and other humans being upset about it.

This principle grounds the meaning of the preferences referred to in the first two principles, but it's not straightforward: for all sorts of reasons, our actions may not perfectly reflect our underlying preferences. I'll say more about this later.

I also want to reinforce the obvious point that we may all have different preferences -all eight billion of us, in all our glorious variety. The machine learns eight billion different predictive models. And I am certainly not proposing to install any particular set of "human values".

Now, unlike Asimov's Laws, these three principles are not laws built into the AI system, that it consults for guidance. They are guides to AI researchers in setting up the formal mathematical problem that their AI system is supposed to solve. And the formal problem should have the following property: if the AI system solves the problem, the results will be provably beneficial to humans.

The kind of problem I and my students have been investigating is called an assistance game. It's a game - a term borrowed from game theory in economics - because there are always two or more decision-making entities involved: at least one we think of as the "human" and at least one who is the "robot". It's an assistance game because the robot's payoff - what it wants to maximise in the game - is the human's payoff, but only the human knows what it is. The robot has a prior belief about the human's payoff, and it can learn more during the game.

Perhaps an example will make this clear: you are the robot; your partner is the human; and you have to buy your partner the perfect birthday present using money from the joint account. You're not sure what to get, and in past years you've usually got it wrong, but your payoff is precisely your partner's happiness with the present.

Now, this particular case is known to be unsolvable, but in all other cases we can set up instances of the assistance game and solve them, which means calculating how the human and the robot should behave. And what we find is exactly what we would hope:

The human has an incentive to teach the robot about their preferences, and the robot defers to the human; it asks permission before carrying out any plan that might violate some unknown preferences.

So, after COP36 asks the AI system to deacidify the oceans, the system asks about our preferences for oxygen before initiating the chemical reaction. And we say, “Yes, thanks for checking, we’d really like to keep it!”

We can also prove in a general sense that the machine will act in a “minimally invasive” way, trying to change only those things it’s already sure we want changed and not messing with the rest. This is really important, because the machine will always have a large amount of uncertainty about our true preferences.

Perhaps the most important result is that the machine will always allow us to switch it off. This is the key to the control problem.

Let’s look at a simple example, first in the classical model: a robot given a fixed objective such as “fetch the coffee”. It thinks to itself,

- I must fetch the coffee
- I can’t fetch the coffee if I’m dead
- Therefore, I must disable my off switch
- And possibly Taser all the other Starbucks customers

Obviously, we don’t want this kind of thinking to happen in the robot, but it seems inevitable if the robot has a fixed objective.

In the new model, the thinking goes as follows:

- The human might switch me off
- But only if I’m doing something wrong
- I don’t know what “wrong” is, but I know I don’t want to do it (that’s the second principle and the first principle)
- Therefore, I should let the human switch me off - because the human will be better off if they decide to do that

We can turn this into a mathematical theorem that links the robot’s incentive to allow itself to be switched off directly to its uncertainty about human preferences. The theorem seems to be robust to all sorts of complications in the basic scenario. So, I think the three principles - particularly the second principle on uncertainty - give us a handle on the core of the control problem.

Now, as you’re listening to my explanation of the three principles and their implications, you’re no doubt thinking of all kinds of difficulties.

Please ask those questions at the end. But to save time, I'll say now, no, machines will not learn to copy evil human behaviour, and no, I'm definitely not ignoring the wellbeing of other animals.

Now, as we move beyond the basic two-player assistance game, we immediately face the question: How should the machine decide, when its actions affect more than one person?

This is a question that moral philosophers and political theorists have studied for thousands of years. I cannot possibly do justice to the literature here because I haven't read it.

I can, however, report that the three major approaches - utilitarianism, virtue ethics, and moral rights - are roughly tied in the philosophy polls (literally!), but virtue ethics and moral rights have formed an alliance to keep the unfashionable utilitarians out of power.

Now obviously, with more than one person, the machine needs to make trade-offs. For example, if everyone wants to be All-Powerful Ruler of the Universe, most people are going to be disappointed. A utilitarian would, roughly speaking, weigh the preferences of everyone equally and maximise their sum. That sounds straightforward enough, and in this case, it might well lead to the sensible conclusion that nobody should be All-Powerful Ruler of the Universe.

But sometimes it's fraught with difficulty, especially when weighing decisions that affect who will exist in the future. In the movie Avengers: Infinity War, for example, Thanos develops and implements the theory that if there were half as many people, everyone who remained would be more than twice as happy. This is the kind of naïve calculation that gives utilitarianism a bad name. The Financial Times' review was a learned disquisition called "Thanos shows us how not to be an economist." Meanwhile on Reddit there was, of course, a subgroup called "Thanosdidnothingwrong," but they kept purging half their members, so it didn't last long.

I bring up this example to make the point that these issues are not "merely" - if that's the right word - philosophical. They really matter, and we must get them right as AI systems approach Thanos levels of power.

Now, there is a school of thought within AI that proposes to avoid trade-offs altogether by building loyal AI systems that serve only their owners' interests. There are all sorts of problems with this approach, as we can see in this conversation between Robbie the robot and Harriet, its human owner:

ROBBIE: Your husband called to remind you about dinner tonight.

HARRIET: Wait! What? What dinner?

ROBBIE: For your twentieth anniversary, at 7 pm.

HARRIET: I can't! I'm meeting the Secretary-General at seven thirty! How did this happen?

ROBBIE: I did warn you, but you overrode my recommendation.

HARRIET: Okay, sorry. But what am I going to do now? I can't just tell the SG I'm too busy!

ROBBIE: Don't worry. Her plane has been delayed - some kind of computer malfunction.

HARRIET: Really? You can do that?!

ROBBIE: The Secretary-General sends her profound apologies and is happy to meet you for lunch tomorrow.

Robbie's solution here is brilliant, but only for Harriet. Loyal Robbies simply won't work - they must consider the preferences of all those they affect.

On the other hand, looking on the bright side, we can see that AI offers a valuable experimental tool for trying out various moral theories in a very literal-minded way.

Now, in addition to dealing with many humans, AI must also deal with real humans, which is particularly relevant to the third principle - how machines learn about human preferences from human behaviour.

As I said earlier, our actions may not perfectly reflect our underlying preferences, so inferring preferences from behaviour is far from easy.

We are myopic, computationally limited, and emotional, leading in many cases to choices we regret.

But the biggest challenge to the three principles is the plasticity of human preferences - the fact that they can change due to external influences.

First, there is a practical problem of ensuring that machines don't mould our preferences to be easier to satisfy - something I think is already happening with social media content selection algorithms.

Second, there is a fundamental, unsolved philosophical problem: if the machine is deciding now to do something for you that will take effect tomorrow, who is it working for? Today's you or tomorrow's you?

Finally, the principles implicitly assume that humans are the autonomous possessors of their own preferences. It's a reasonable starting point, but it's not valid in the long run.

Amartya Sen, the great economist and philosopher, emphasised that an oppressive society moulds the preferences of individuals so that they accept or even welcome their oppression; therefore, it may not be appropriate to take the preferences at face value. So, should AI systems be in the business of moulding human preferences in "better" directions, whatever that means? Possibly, but that's a place where even the angels fear to tread.

Now, I'm often asked whether we're ready to encode the three principles into legislation and detailed regulations. No, not yet. We need a lot more theoretical and experimental work before we have the necessary design templates that could form the basis for regulation.

Fortunately, I believe companies will have a very strong economic incentive to adopt this new approach as soon as it's feasible.

For example, suppose you have a domestic robot built according to the classical model with fixed but imperfect objectives.

And you're stuck at work late, your partner is away, perhaps looking for a birthday present, and the robot is looking after the kids for you. Now the kids are hungry and very grumpy, and there's nothing in the fridge, and there's not time to go shopping.

And then…the robot sees the cat.

Unfortunately, the robot lacks the understanding that the cat's sentimental value is far more important than its nutritional value.

So, well, anyway, you can imagine what happens next.

And then the newspapers find out and go bananas, and that's the end of the domestic robot industry, because no one would ever buy a robot that might do such a thing.

So having this kind of humility - knowing that it doesn't know all of our preferences and asking before doing something rash - is going to be an economic necessity for human-facing applications of AI.

Stepping back a little, I think we have to move from the current situation, where AI researchers think they're doing good AI, but the ethicists are wagging their fingers and saying "Bad, bad!" to a situation where the AI researcher gets up in the morning and doesn't just say "Okay, okay, I'm going to listen to those insufferable ethicists today," but instead, says, "Today I'm going to build a really high-quality AI system."

And what that means is an AI system that's provably beneficial to humans, just as when a doctor strives to be a good doctor, what that means is healing people and not lining one's pockets by selling fake medicine.

To close my lecture, I'd like to bring up the nature of our co-existence with AI, assuming we have solved the control problem and developed general-purpose, provably beneficial AI.

One possibility is that an increasing dependence on AI leads us to become enfeebled and infantilised, like the humans in the film WALL-E.

But before WALL-E, there was E M Forster's The Machine Stops, published in 1909. The Machine of the title is an all-encompassing intelligent infrastructure that meets all human needs. Forster depicts the internet, email, email backlogs, videoconferencing, iPads, massive open online courses or MOOCs, widespread obesity, agoraphobia, and avoidance of face-to-face contact. Humans become increasingly dependent on the Machine, but they understand less and less about how it works. Kuno, the main character, sees what is unfolding but is powerless to stop it:

"Cannot you see, cannot all you lecturers see, that it is we that are dying, and that down here the only thing that really lives is the Machine? We created the Machine to do our will, but we cannot make it do our will now. It has robbed us of the sense of space and of the sense of touch, it has blurred every human relation, it has paralysed our bodies and our wills…Oh, I have no remedy - or, at least, only one - to tell men again and again that I have seen the hills of Wessex as Aelfrid saw them when he overthrew the Danes."

The first lesson of Forster's story is that as we gradually hand over the management of our civilisation to machines, we lose the ability to do it ourselves, and the next generation loses the incentive to learn how to do it, and the chain breaks. Since the dawn of humanity, we have spent roughly a trillion person-years just passing on what we know to the next generation, through thousands of generations, to keep our civilisation alive and growing. What happens when none of that is necessary?

But there's perhaps an even more important lesson: What Kuno feels, when he escapes the safe confines of the Machine, reaches the uninhabitable surface, and sees the hills of Wessex, what he feels is autonomy: the counterfactual freedom to deviate from the path that was prepared for him, the path that he'd prefer.

Autonomy is a fundamental human value, which means that beneficial AI systems cannot ensure the best possible future if ensuring means a loss of autonomy for humans. It may be that machines must refrain from using their powers to predict how we will behave, in order for us to retain the necessary illusion of free will.

However we resolve this self-referential puzzle, our AI systems must and will learn to stand back, as parents do, eventually to say, "No, I'm not tying your shoelaces, today. You must do it yourself." They will not create the WALL-E world unless we force them to.

But the parent-child relationship is not the right metaphor, because we (the children) will have all the power, even though the machines will in fact be far more powerful. We need a new metaphor, a new way of seeing ourselves, and we will need all the writers and filmmakers and poets to guide our culture in the process.

Thank you.

(AUDIENCE APPLAUSE)

ANITA ANAND: Again, a completely fascinating lecture. Throughout this series you have managed to scare me about most things. I mean, from sort of Armageddon to eating a cat. You've given these ways that we might be able to stop it, but don't we need everybody to agree that this is how we're going to stop it, and at the moment we have a world where people can agree on nothing, where governments can't agree on imminent threats like climate change or nuclear disarmament or how to feed the poor. How do you – or do you – expect they can

agree these three principles that will stop us eating pets and annihilating ourselves?

STUART RUSSELL: I think there are a few things working in our favour. One is that if we do achieve general-purpose AI, it will be such an immense generator of wealth that trying to hog it to yourself serves no purpose whatsoever. It will be like hogging digital copies of the newspaper; simply won't make sense. Another reason is that it's in no-one's interest if someone makes a mistake and creates general-purpose AI that is uncontrollable, just as it's in no-one's interest for nuclear power stations to explode, as happened with Chernobyl and Fukushima.

So, we actually got our act together reasonably well after World War II to make sure that nuclear energy was safe. It didn't quite work but we did actually cooperate fairly well. So, I think that the major nations in the world will cooperate to try to develop safe AI and I am actively encouraging connections between the US and China, Britain, Russia and so on, to make this happen.

ANITA ANAND: Well, that's really interesting because you're in the room where it happens and how freely do they exchange ideas?

STUART RUSSELL: On this topic very freely because safety doesn't instantly confer a military advantage or anything like that, and my discussions in China, for example, people are very open to this idea.

ANITA ANAND: And we're going to open this up to this splendid audience here at the National Innovation Centre for Data in Newcastle, but just one small one from me. You've asked profound questions that involve, in your words, "annoying ethicists" and, you know, philosophy - what do humans say they want and what do they really want. To me, it seems like every AI centre that's working ought to have a resident philosopher or ethicist on board to prick the consciences or at least remind people that this is important. Is that happening at all? Are those people there?

STUART RUSSELL: Yes, they are, but there is a bit of finger wagging going on and I think that's unproductive. It doesn't work for the ethicist to be leaning over the shoulder of the AI researcher saying, "Bad, bad." What works is for a real conversation to happen, for the AI researchers to understand that they don't know much about the last two and a-half thousand years of ethics research and be willing to learn about it and read it, and just see the pitfalls because philosophers, in a way, have been debugging the moral programmes of other philosophers. Some philosophers say, "Well, you should do this, this and this. This

the principles we should all follow," and another philosopher says, "Well, if we did that, then this terrible thing would happen, and we'd all die."

So, in that debugging process they have developed very finely-honed skills of spotting flaws in overly general principles that wouldn't actually work, and we could really benefit in AI from that experience. So, the example that I gave of changing the size of the population, getting rid of half the people, actually comes from a 19th century philosopher, Sidgwick, who actually proposed a solution pretty similar to Thanos's.

ANITA ANAND: Gosh. Let's open this up. If you could wait for the microphone, let me remind you, and say who you are. Right, there's one question over there. Let's go there first.

ALEX FAWCETT: Hello. I'm Alex Fawcett, Ecosystem Director at Sage, and I've got a question about trust. So, do you think there's already an issue with public trust of AI and what can companies who are building AI, like Sage, do to regain it?

STUART RUSSELL: I believe there is a big problem of public trust. For example, in the area of self-driving cars, trust has dropped precipitously. From the high seventies, I think, people say they would be willing, down to the 30 per cent or something like that. So, part of it actually has to do with the fact that the technology turns out not to be as robust as some people would like to claim.

Deep learning is actually quite difficult to get right, and often it appears to be working but you make a slight change in the circumstances - for example, the algorithms that have learned to recognise cancerous skin lesions, turn out to completely fail if you rotate the photograph by 45 degrees. That doesn't instil a lot of confidence in this technology.

So, I think there's two parts to this. One is, develop better, more robust methodologies, don't overclaim, but also explain to the public how it works, where it's appropriate, and especially where it's not appropriate to be used. I think people are very afraid of AI being used to do things like keep their kids out of university or refuse their job application, where they really shouldn't be using algorithms.

ANITA ANAND: Thank you very much. The woman on the end of the line there.

KIRSTEN RICHARDSON: Hi, Kirsten Richardson, a PhD student from the School of Computer Science. I was thinking about nature and the climate crisis

being a good example of really intricate relationships and interdependencies that we don't understand, and in the assistance-game experiments you were talking about the humans teaching the machine about our preferences; is there scope for the machine teaching us about when our preferences might be wrong?

STUART RUSSELL: As I said, moulding human preferences in better directions is where the angels fear to tread because you can only imagine how badly wrong that could go, and politicians do this to us all the time and we don't like it. So, philosophically, the whole notion of what is a better preference, telling someone that actually they're wrong about which future life they prefer is a really difficult thing to do.

Another connection to your question, climate is a really interesting case because one could argue, and in fact some people have written articles saying this, that we don't need to wait 20 or 30 years to see what happens with a super-intelligent machine, that corporations function as machines. They optimise and misspecified objective which is, let's say, quarterly profit, ignoring the externalities, ignoring all the problems that they cause for the rest of the world, and the fossil fuel industry has outwitted the human race, right. We have lost. I'm sorry. We have lost. Even though we all know what needs to be done, we have lost because they figured this out 50 years ago and have developed a strategy that has outwitted the rest of us.

So, we can look at that example and say, "If you want to see what uncontrolled super-intelligent AI is like, it's like that except worse."

ANITA ANAND: Gentleman in the stripey shirt?

LEON DRISCOLL: I'm Leon Driscoll. I'm a GCSE student here in Newcastle. You said that there's a disincentive for AI companies to develop AI unsafely, but shouldn't be the opposite be true as well? If a quick, cheap and unsafe development of AI allows a company or other organisation to gain a first mover advantage from the development of artificial intelligence, surely, they have a strong incentive to develop AI quickly and unsafely?

STUART RUSSELL: Yes. This argument, sometimes called the "racing argument," is something that worries people, particularly between nations - if one nation wants to try to get a lead, they might cut corners. And in other areas this is why we develop these regulatory bodies, consortia sometimes. So, for example, with electricity, the various electricity providers and developers of appliances got together and said, "We're not going to gain acceptance until we face up to the safety problems." There were lots of fires, there were lots of electrocutions with early electrical devices and wiring, and so they developed

standards. And if you look on your plugs and [toasters] and so on, there's little marks that actually come from the various standard institutes and you can't sell those appliances without the mark, they have to meet those standards. And that's partly industry regulation and partly legal standards, and it varies by country, but it's fairly successful.

And the partnership on AI, which is a consortium of all the major tech companies, except for some of the Chinese ones, actually is trying to develop these codes of conduct and so on, but what's missing, really, is well, what should the standard be, and that's on the AI researchers to develop the standard. So, we're trying to do that, for example, with face recognition avoiding algorithmic bias, and I think we're fairly close on having technical standards for how to do that properly.

But on the question of the long-term safety of general-purpose AI systems we're, as I said in my talks, still some way away from knowing exactly how to define what the algorithm templates should be. And it's no good saying to Google and Facebook, "You have to do this," if we don't know how to do it. So, we have to solve those technical problems.

EMILY MILES: Emily Miles, Chief Executive of the Food Standards Agency, so a regulator, but I'm not going to ask about that. I was interested in your principle about understanding human preferences and I know as the food regulator that we privilege the short term over the long term. So, it's much easier for us at the FSA to withdraw product that is going to make you sick now than it is to intervene on food that might make you sick in the long term because it makes you fat, for example. So, there's a risk that the artificial intelligence amplifies the human preferences for now rather than later or even for future generations. Is that a problem?

STUART RUSSELL: I think, in principle, no, it won't happen that way because the AI system recognises that we behave in ways that violate our true preferences. So, if you were watching this movie that depicts your whole future life and you saw that from the age of 50 onwards you were obese and possibly even suffering knee problems and heart problems and everything else as a result, you would actually say, "No, I don't want that life." So, your true underlying preferences are not to become extremely unhealthy as a result of eating this piece of cake, it's just that your actual decisions that get made suffer from this myopic bias.

And working with some cognitive scientists and psychologists, we've actually been able to develop methods where the AI system helps humans to

bring the future into the present so that they can actually make decisions that are closer to their own long-term interest.

ANITA ANAND: Thank you. Let's take the question here?

PAUL WATSON: Paul Watson from the National Innovation Centre for Data at Newcastle University. You mainly focused your lecture on individuals, but we structure society in terms of organisations, be they countries or companies or universities, so I was wondering how you felt that the right of AI will affect organisations?

STUART RUSSELL: I think one of the things that AI could do for us, and this is not yet a very well-developed subfield, is improve coordination. For example, we could all agree to cut our greenhouse gas emissions and have a future. There's a coordination failure, right. Everyone says, "Well, I'm not going to do it until they do it," so no-one does it, and this is what game theorists call "a prisoner's dilemma." You rat on your accomplice because that way you get off, but then you both rat on each other and then you both go to prison for ever.

So, prisoner's dilemma is precisely this. There is a better solution, which is that neither of rats on the other one, but you can't reach it without some coordination, right, you have to somehow know or trust or previously agree, and AI systems can help this coordination process by making it clear by generating incentives and finding ways of doing escrow and other kinds of agreements that make these things possible.

So, it's a really interesting question and I've recently, actually, been working in what's called "team theory," which is the game theoretic analysis of how organisations, all of whose members are working towards the same goal, but they're all disconnected from each other, how can that be successful?

LILLIAN EDWARDS: Hi. I'm Lillian Edwards, Professor of Law, Innovation and Society at Newcastle Law School, which is a fancy term for Professor of Technology [Law]. If we're going to develop, or if general-purpose AI is going to emerge, shouldn't we start now with regulating the people who are developing these systems, right, because they're not going to spontaneously [meta morph], right?

And secondly, you may be aware, that in fact there is a proposal on the decks in the EU for the comprehensive regulation of AI and without going into any of the details, I think the interesting point which hasn't been mentioned, is that it isn't based on people's preferences, it's based on human rights, civil and political rights, to things like fairness, equality and non-discrimination, and I

wonder if this isn't – forgive me – a very consumerist view of what we want from strong AI?

ANITA ANAND: Thank you very much.

STUART RUSSELL: So, as I mentioned, right, there is this sort of three-way tie between rights-based, virtue-based and utilitarian approaches to ethics, and my personal belief, and this is a long argument that we're not going to get into tonight, but my personal belief is that the rights-based approach actually can be derived from the utilitarian or preference-based approach but it's a complicated derivation.

But the idea of regulating the researchers now is an interesting one. I could certainly see that we might start requiring much more ethics training, which in other engineering fields has been taken for granted for a hundred years or so, at least. So, to be a professional engineer, whether it's a civil engineer or a mechanical engineer, in the US you have to have ethics training.

ANITA ANAND: But this is about more than training. This is about regulation, this is about someone saying, "You will not go there. You will not do this."

STUART RUSSELL: The regulations that come with the EU law, which I spent a great deal of time trying to fix - actually, just a little anecdote, right, at one point the EU Parliament debated whether Asimov's Three Laws, which were devised by Isaac Asimov to produce interesting storylines for science fiction, they were debating whether to actually enshrine those in EU law. Fortunately, we nipped that one in the bud, and the laws as they currently are proposed have a number of important things, like banning the use of AI to impersonate human beings, which I argued very strongly for because I think it has all sorts of problems and really no legitimate uses except in maybe certain kinds of psychiatric dialogue and so on, but those could be carved out.

So, that's a regulation not on – at least I don't think of it as a regulation on AI researchers, I think it's a regulation on products. So, the EU law is mostly about products, not saying you, the AI researcher, cannot do research on this, "You cannot write that algorithm." They're saying, "You can write the algorithm, but you can't sell it as a product," and that's how it's regulated.

You're right, they talk about two things. They want to regulate high-risk systems, and I think initially most people thought, "Oh, that's just self-driving cars that could kill you or kill a pedestrian, or some medical device that's going to fry you with radiation or something like that," but possibly very cleverly, a high-

risk system is something that can impinge on fundamental human rights and in the EU Charter of Fundamental Rights - not in the Universal Declaration but in the EU Charter – there's a right to mental integrity, and that's really important because that means that any human-facing information system, including all the social media algorithms, could be a risk to mental integrity and therefore is subject to this regulation. So, I think that's very important.

ANITA ANAND: Thank you. I'm going to take a few more questions. Now, I can see some hands over there but sometimes I see somebody in the audience who I recognise and who is interesting. We have Tom Kirkwood with us, who is a Reith Lecturer from the past, and a professor of Medicine, and gerontology is your specialism. How many years ago did you give the Reith Lecture?

TOM KIRKWOOD: Twenty years ago.

ANITA ANAND: Twenty years ago. So, from what you've heard today, I mean, does it raise any questions in your own field?

TOM KIRKWOOD: Well, a big question for me is the question of time. I think at the moment we see the pace at which the human future unfolds is governed by the relatively sedate way in which our discoveries, insights and fashions go forward. Now, with AI, the speed of change could potentially become very much faster.

So, the question I have is, do you foresee issues down the line in reconciling the very different rates at which AI and human futures might play out?

STUART RUSSELL: Absolutely, yes. I think we need to start preparing now. In fact, we're already in a period of pretty rapid change. I mean, when I think about the 70 years or so we've been doing AI, just in the last decade we've knocked over agile leg locomotion, recognition of objects in images, speech recognition, machine translation. These are major open problems for 70 years and now they're solved. And the level of investment, I would guess in the last five years, more has been invested in AI than the previous 65 years put together. So, we can only expect that things will accelerate.

My goal here is to get people to start thinking about these issues now and not find ourselves caught short when the next big step happens and we're not ready for it, and all kinds of mayhem happens, as I think has happened with social media algorithms, and yet, because the algorithms were making tons of money, we haven't been able to switch them off.

ANITA ANAND: Thank you. You've been waiting patiently and patiently, so let's go over there?

POLLUM. I'm Pollum, a computer scientist in Newcastle University. So, you've been talking about our preferences and, basically, that AI et cetera is serving our needs as if there was a clear, well-defined boundary between us. If all goes well, there's that wonderful AI kind and then humankind, and one could argue that actually AI kind is actually better than us (48:39) and so forth. Will it ever be the case that AI will develop their own preferences? In other words, if you blur this line, what happens? Then if you project into the future, then will it not be the case that the AI are also entitled to express their preferences and then the whole picture – what happens then?

STUART RUSSELL: If we discover that, probably entirely by accident, we've created conscious machines that have real subjective experience and suffer, for example, having to listen to us, then that does change the ballgame. It changes the rules completely. But the fact is we have absolutely no clue, despite periodic, every 20 years I suppose, roughly every generation philosophers get excited about consciousness and say, "Perhaps we can really nail the problem," but they never do.

We're just left with the fact that we have no way to create consciousness. No-one I know is seriously working on that problem. No way of detecting consciousness - even in human beings. We can tell that there's nervous activity but that's no different from me taking my cell phone and saying, "Yeah, there's electrical activity in my cell phone." Is it conscious? No idea, right, and so, I have to just leave that problem for future generations because, like anybody else, I have no answer to those questions.

ANITA ANAND: And sometimes that's okay. And a question from behind there?

RACHEL FRANKLIN: Hi. Rachel Franklin, I'm a Professor of Geography here at Newcastle University. I suppose my question has to do with preferences and the negotiation of preferences, and I think a lot of the examples that we've seen tonight have been at sort of the macroscale, right, how institutions negotiate preferences or countries, or cars and drivers, but I'm really curious about the microscale. The household scale, I think, is maybe where I'm most curious about the potential for conflict and the question whether machines can do better humans, and the negotiation of preferences, for example, between husbands and wives.

It's something that we can't figure out as humans, how is it that you see AI negotiating that very basic level of conflict and how you could see AI doing better than maybe the typical householders manage to do?

STUART RUSSELL: I'm not really a family therapist but I understand that family therapists are sometimes really helpful simply by asking people to express their preferences, their annoyances. My feeling is that human beings are basically good, they want to live in harmony with each other, they love each other. But one of the things I said in the third lecture, which was about the future of work, is that this is an area where really humans have a comparative advantage because we've been there.

So, if there is a future of work, if we're not all just going to be lotus eaters, it's going to be in this interpersonal role where we have this enormous comparative advantage, and we will become extremely skilled family therapists, who perhaps will use AI tools to scan tons and tons of data about individuals and learn all their personal mechanics, so to speak, but this will be our future role is to make each other's lives better by direct intervention, so to speak.

ANITA ANAND: Let's try and squeeze in one cheeky, final question?

HERB KIM: Hi, my name's Herb Kim. I'm the founder of Thinking Digital and TEDx Newcastle here locally. My question is, I mean, you said something to me that was actually quite remarkable, that if someone discovered, engineered general AI, that there would be this natural desire or willingness to share it amongst others, partially because of the amount of wealth that was being generated, but I was thinking more from a perspective, I guess, of national defence and I was thinking if President Xi's scientists were to come up to him and say, "You know folks, boss, we've cracked it. We've done it here in China. No-one else knows about it at the moment. Do you want us to share it or would you like to kind of keep a handle on it for a while?" I mean, you can take your guess as to how he might react to that.

I'm just wondering, sitting here, are we actually taking this seriously enough, should President Biden be advised to cut defence spending by 20 per cent and put that money into helping to engineer general AI, as well as also just planning out what on earth we're going to do if this thing actually shows up so that there's actually a plan that as we approach that threshold that it won't just suddenly effectively overwhelm us, which clearly [throughout] the day that would happen?

ANITA ANAND: Thank you.

STUART RUSSELL: Some sceptics say, you know, if you go and talk to the real AI researchers, nobody is building summoning chambers for these demons that we might create, but actually, the National Science Foundation now has in its plan for AI funding explicit coverage for safety and control research and, oddly enough, Xi Jinping is the only world leader who has explicitly acknowledged that AI could be an existential threat to humanity. So, I think there's official guidance from the top that we need to solve this problem in China.

You bring up the question of defence. I actually think that the defence problem is quite urgent. I devoted the second lecture to this topic, and it doesn't require anything close to general-purpose AI. In fact, it's probably quite a bit easier than a self-driving car to build an extremely effective and dangerous autonomous weapon, and you can actually buy it now.

ANITA ANAND: Do you sleep at night?

STUART RUSSELL: Less so, but I think that's just as I'm getting older. I feel reasonably optimistic, actually, on the long-term question of will we be able to control our creations because I think we will get our act together. I'm just making one initial proposal. There are other proposals out there and as we start to experiment, we'll try these things out in simulated worlds and we'll see, "Oh, look, it doesn't quite work because this thing goes wrong and that thing goes wrong," and we will get there, I think, to have safe AI systems.

But the question of how society interacts with these new technologies, these are really hard questions. I'm glad to say that my colleagues at Berkeley, who are in the social sciences and humanities, this is what many of them want to do – not all of them, some of them still want to study Victorian poetry, and that's great – but many of them actually want to leave their home departments, move into the College of Engineering, which is sort of like the dark side for them, and actually help us figure out how to navigate the next 30 years safely for the whole human race.

ANITA ANAND: We're going to have to leave it there. Thank you so much, Stuart, what a thought-provoking lecture, and a big thanks to our hosts here at the National Innovation Centre for Data in the University of Newcastle. That is it for the Reith Lectures for another year. Stuart's series on AI and the huge Reith Archive is available via the website. Do please check those out.

But for now, from Newcastle, goodbye.

(AUDIENCE APPLAUSE)