---

# 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)