
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)