All Party Parliamentary Group on Drones

Oral evidence: APPG Inquiry into '[The Use of Armed Drones: Working with Partners]'

Tuesday 31 October 2017

Witnesses: Professor Stuart Russell, Dr Tom Simpson

Inquiry Members present: Clive Lewis MP; Baroness Stern; Lord MacDonald QC

Chair: Professor Michael Clarke

Q1 Lord MacDonald: Can we just start with something fairly broad? Professor Russell, I wonder whether you could address this in the first instance, and that's really to tell us what you consider to be the main legal and ethical concerns that emerge from the development of lethal autonomous weapon systems. So just your headline points on that, really.

> **Stuart Russell**: So, the issue has been in the international community since around 2012 with the Heynes Report. The main point of discussion has been compliance with IHL, the ability to discriminate combatants and non-combatants, the ability to establish proportionality in assessments. I think these are significant issues, but to me they miss the main point. I think the mental picture that people have is that right now we have drones that are piloted by humans, and they have certain characteristics. What happens if you replace the human pilot by an AI system? I think that's the wrong mental picture altogether, because when you have autonomy you can use the weapons in completely different ways, and in particular, the notion of autonomy that the United Nations is using, which is the ability to locate and engage targets without any human intervention, means that you can make scalable weapons. So 'scalable' in the IT world means that you can deploy as many servers as you want, you can build giant server farms, in this case you could deploy as many weapons as you want with a fixed number of people to define a mission and deploy the weapons. So, in that sense you have weapons of mass destruction, you can launch attacks to wipe out an entire country if you so wish, and if you can afford the weapons. You don't need a large military industrial complex or infrastructure to do this. So to me, and I think to many in the AI community, that's the primary concern. It creates an extremely destabilising situation because you have difficulties of attributions, so weapons could be untraceable, and you expose yourself to no risk by using them.

Lord MacDonald: Human Rights Watch, I think, has suggested that lethal autonomous weapons could never be compatible with international humanitarian law, and also that they, in the view of Human Rights Watch, would increase the risk of human or collateral casualties. Dealing first with international humanitarian law, do you agree with them about that?

Stuart Russell: I think that's a moving target. I don't think you could say they would never be as capable of discrimination as a human soldier. They would make errors, they would be susceptible to counter measures whereby people would figure out how to disguise themselves. It used to be a joke because the ability of AI systems to recognise objects has now exceeded the human capabilities for natural occurring images. You could construct images that are designed to fool the AI system. Those images are quite robust. We used to joke that you could wear a t-shirt that would cause the system to think you were a penguin, rather than a human being, but in fact these images are fairly robust. You can take a cell phone picture of the image, you can print it out, you can take another cell phone picture of the image, you can look at it from a different angle and it will still fool the AI system into thinking it's a penguin. So I would imagine that combatants would probably develop various kinds of counter measures that would make them more difficult for the system to detect. That aside, I suspect that the discrimination capabilities are either comparable to those in the human or will soon be comparable to those in the human.

The thing that AI systems are unlikely to be able to do is to judge proportionality and military necessity. So whether that's relevant depends on the scope of the mission that you give to the machines. If you give them a mission where you already know that within a certain geographical region, all the combatants are actively engaged, they haven't surrendered and so on, then military necessity is in the judgement of the human, and the mission can simply be, you know, engage and eliminate all the combatants you find in some geographical region. The machine doesn't have to make a judgement on military necessity, as long as the human has a reasonable capability to decide in advance that all such engagements will be compliant with IHL. If you had a larger mission, such as 'defend the border against any incursions', then you're handing over much more responsibility to the machine, and you wouldn't be able to guarantee any engagement would be compliant.

Lord MacDonald: Why shouldn't you be able to develop systems that were capable of making judgements about proportionality and necessity, through observation, for example?

Stuart Russell: In principle, eventually, you could. It's not something that's likely to be feasible, I'd say, within the next decade or so. It requires an understanding of the situation, the military situation. For example, in the first Gulf War, my understanding is that it was judged militarily unnecessary to attack the retreating Iraqi armoured columns, and so they let them all go back to Iraq without attacking them. That was a judgement call made, presumably, by the highest political level. It's not something that I would entrust with a machine in the near future.

Lord MacDonald: What about the broader ethical debate about the implications of delegating life and death decisions to machines? I think Germany has said, hasn't it,

it would never go down the path of utilising machines that were effectively making those kinds of decisions, and they would always want those sorts of decisions to be made by a human. I think some other states have taken a similar view. What's your stand on that?

> **Stuart Russell**: So my personal stand is that it would have a serious effect on the perception of the dignity of human life, that if machines are the ones who can decide, essentially, who gets to live and who gets to die -, but I don't imagine that this argument is going to be convincing to the major powers who are concerned about strategic parity and security and those kinds of issues. I think if you wanted to come up with a rule, a very simple rule for an ethical code, let's say for the professional societies, a rule saying don't make machines that can decide to kill human beings is a very straight forward rule. As a commandment, it could carry a great deal of weight. The public opinion, the moral stigma and the simplicity of that message, I think all carry a lot of weight for me, but I understand that they wouldn't necessarily carry weight with the Chinese government and their military calculations.

Q2. Lord MacDonald: Dr Simpson, I'd be interested in your view as an ethicist, of this issue of machines making life and death decisions over humans, and also any other aspects of the discussion which have come out so far, if you'd like to comment on them.

> **Tom Simpson**: So just in terms of the structure of the debate, my expertise is as a philosopher, so I'll defer to Stuart on the technology. The philosophical issue, which is behind the legal issue is the question of who holds responsibility for when an automated system kills someone wrongfully. There are arguments that, in principle, no person that could be held responsible for that, and therefore to use these systems would contravene laws of war. I'm pretty unpersuaded by those arguments, so it seems to me just very clear that those public bodies which authorise automated weapon systems for use are responsible for the unintended consequences that come about as a result of, in particular, innocent people who die as a result of the decisions the machines make. It seems to me that the question is really, is it justified to impose that risk on innocent populations? So my own view in this context is that, so long as machines meet a threshold where they're imposing less risk than a normal human combat force, then actually that's an entirely justified imposition of risk.
>
> The person who's exposed to that risk is basically faced with the choice, do you want more chance to be killed by human combat force, versus less chance of being killed by a robot? It seems to me that the rational thing in that situation is to say, 'I'll take the lower risk.' This goes to the deep question of whether there's some kind of fundamental opposition to the principle of machines being in power to take human lives. One question is explaining why that impulse exists, and then another question is how justified it is. In terms of explanation, there's this fascinating book, Barbara Ehrenreich on Blood Rites, she looks back to our evolutionary 'just so' past, and says one of our most

urgent human fears is predation. So if you look at a lot of Hollywood horror films, you know, Jaws, Arachnophobia, what's the one with Samuel Jackson on an aeroplane with snakes? It's about animals hunting humans, and a loss of control. I think robots trump this atavistic fear of interspecies predation, with a predator that we can't control. So be it almost like a species disloyalty that we create. So, the instinct that robots prime are very deep, and actually might be so strong to make it rational to say, 'Actually, we'd rather take a high risk of being killed by people as opposed to a low risk of being killed by machines.'

Lord MacDonald: Can I just pick up one thing which you just said, which was that someone who is the object of the risk might take the lower risk, even if it means being killed by a robot? That's a survival choice, isn't it? Which doesn't necessarily make it the right policy on the part of the individual who's deciding whether to use a robot or a human being.

Tom Simpson: Yes. So obviously a question of, nobody's giving consent in this scenario, so we don't have actual consent. The question is, could we presume you would give consent were you to be forced on that choice? I suppose my sense is that we will actually see the moral stigma around robot killings reduce as it happens more. So it's a trivial example but it is an example, is automated checkouts at Sainsbury's. So when these came in, you can still google the news stories that came out in 2008 and 2009 about the loss of human contact, the value about paying for your shopping with a person, as opposed to a machine. Now it's just quicker and easier to do it and you just go through. The efficiency considerations have this dominating influence, and this power to change norms over time. I think the fifteen-year prospect might be more sanguine.

Prof Clarke: Yes, Baroness Stern, if you please.

Baroness Stern: I was thinking about Sainsbury's, really.

Prof Clarke: Is it easier? No.

Vivien Stern: You're right in the sense that I wouldn't go near them, but I don't think it's because all of the things about philosophy, I think it's because it's putting people out of work. Just take me on the journey from people have got used to those checkouts at Sainsbury's, and so they'll get used to somebody next to them being killed by a robot. It's a bit of a journey you're taking us on, I wonder if you could just put steps in.

Tom Simpson: The history of military technology is trying to make it easier for us to kill the other side without being killed ourselves.

Baroness Stern: Is that right? The history of military technology?

Tom Simpson: Well, say bows and arrows are a way of stepping back from a distance, artillery is stepping back, the rifle (inaudible 15.54).

Lord MacDonald: Shield formats, Roman shield formations, yes.

Tom Simpson: Every step along the way, what you get is this discourse. The British army were on the wrong side of this in the American Revolutionary wars in the late eighteenth century because they were in camouflage, shooting at us, and they weren't standing up and fighting. So there are allegations of cowardliness and not fighting properly, but in the end the technological advantages win out, and then you recalibrate your perception of the combat. It's just, this is a fair fight and this is the way we engage with each other. It just seems to be that with automated weapon systems, the same recalibration will take place, and whereas one side possessed artillery and the other side didn't, there's a sense of unfairness that you can be killed without any reply. We just come to accept that that's part of the dynamic of war.

Baroness Stern: Nuclear weapons?

Tom Simpson: Right, so this is actually very helpful because one of the questions that we face here in the long terms is the interplay between some morally desirable norms, versus what is practically feasible. So although I don't have an issue in and of itself of robots being in power to take the decision to kill humans, there may be all sorts of unintended consequences from stepping off the path of developing these systems and deploying them. So there may well be an argument that if we can make it stick a no development norm internationally, which is what the current convention CCW process is, that has been advocated in that process. If we can make that stick, I can see that being an attractive possibility. The question is, can we make it stick? in particular, are there great power adversaries, if not now, at some point potentially in the future who would not abide by that law? So what does seem to be a viable norm is no first use. So what we have with nuclear is, the technology exists, you can't go back on it, but we had a no-first-use policy, and that's a plausible sticking point.

Q3. Prof Clarke: Yes, this is an arms control approach. Just before we carry on with Baroness Stern's line of questioning, can I just go back on something to make sure I understand this, because I think it's a really important point that Professor Russell said early on. I just want to make sure I've got this right. You know, where you're saying that law, the lethal autonomous weapons are a journey into, in effect, weapons of mass destruction. Yes? Which is a very important point, because what you're saying is that they can create an environment which may have the same effect as a WMD, of flooding an area with very small weapons, very small devices that can kill a lot of people. Am I right in understanding that?

Stuart Russell: Yes, that's one example. I did some unpleasant research on the web and found out that 1g of high explosive, if it's properly shaped, can

blow a hole in 1cm of steel. 1g of high explosive can easily be carried by a 1in in diameter helicopter, which can also carry the kind of chip that Mobileye produces for the self-driving cars. So you can build a lethal weapon that weighs about 11g, and you could fit 3 million of those in a truck. You can programme them, so unlike nuclear chemical weapons, you can decide, 'Okay, we're only going to kill males between the age of twelve and 50,' for example. Over 50s. They are also unlike nuclear weapons, where you really have to cross a big threshold to decide that you're going to use them. You can dial the amount of deployment any way you want, you can say, 'I only want 25,612 deaths today,' and that's exactly what you get.

Lord MacDonald: You spoke before about-, not here, but I remember about you speaking directly about the real possibility in the not too distant future or being able to send hundreds, if not thousands, of these tiny little weapons into areas where there are a lot of human beings, and killing very large numbers of people who are selected in the way that you've just described, by a physical characteristic, by clothing even.

Stuart Russell: Or by ethnicity.

Lord MacDonald: Or by ethnicity, yes.

Prof Clarke: I'm just trying to think this through. Is it the autonomy of the system that bothers you, or the miniaturisation and discrimination? Presumably that can be done via a weapons system which is just programmed, but not particularly autonomous. I'm trying to work out whether it's the autonomous nature of the system which makes it a gateway to WMD.

Stuart Russell: Yes, I'm using the UN definition of autonomy. It's precisely that capability which allows it to be scalable. You don't need the human to be guiding the weapon, or deciding that's a target, deciding where it will go. When I hear that the whole notion of autonomy is so ill-defined and we don't really know what we're talking about and it may never exist, this is pure obfuscation. The type of autonomy that causes scalability is the same type of autonomy that chess programmes have. They decide where to move the pieces, they decide which enemy pieces to kill, and they do that using algorithms that have been around for, arguably, 105 years. So this is not some magical thing that requires philosophers to work out what exactly we mean by autonomy, it's a pretty straight-forward technological concept that already exists. When you look at a self-driving car, you say, 'Take me to the airport,' there may be a very high number of routes to get from here to Heathrow, and it will choose one based on current circumstances and traffic and all the rest, and will execute it, maybe change its mind half way through. That's the kind of autonomy that would lead to scalability.

I think it's really important too that we understand the difference. We talk about, you know, for example, machine guns, Kalashnikovs. Sure, with 3 million Kalashnikovs you can do an awful lot of damage, but only if you have 3

million soldiers to carry them, which means a support structure of another 15 million people to train and transport and fuel those soldiers. So you're talking about a nation state. Nation states are subject to all the checks and balances of the international system, and sometimes those fail but not very often. If you want to launch 3 million autonomous weapons, you just need-, I don't know how much they're going to cost, maybe $50 each. You just need $150 million and a couple of programmers to set up the mission definition, and that's it. You don't need a nation state. So I think they would be much more destabilising in that sense, than even nuclear weapons.

Q4. Prof Clarke: Okay. So, Doctor Simpson, you were just getting onto the idea that maybe an arms control approach is feasible. Do you accept the notion that autonomous systems, lethal autonomous weapons could have the same effect as WMD, could in effect have to be dealt with as if they had the same moral and military equivalence of nuclear weapons? Do you go along with that?

Tom Simpson: Yes, certainly. The levels of destruction that Professor Russell has described are absolutely on par with WMD. One of the tricks here, which I don't have any deep insight into, but there are two different directions in the development of computer technology, so there's very high end stuff which is the fighter planes of the future, and then there's the low end stuff. Actually, the arms control regime for the high end stuff is, in principle, relatively easy to manage, along the same lines that we've already got. It's the risk of bad actors, individuals, terrorist groups.

Lord MacDonald: Non-state actors?

Tom Simpson: Non-state actor getting their hands on the low-end stuff, which poses the real threat. I guess the pertinent question is how you then defend yourself against those technologies.

Lord MacDonald: Those sorts of technologies that you've been describing, we're not there yet but is it likely that within ten, fifteen years or so, that sort of technology would be available to the non-state act that Tom was talking about? You need the hardware, don't you? You need the actual devices to send off into the ether?

Stuart Russell: Yes. They would be, in some sense, as difficult to control as small arms.

Prof Clarke: Handguns and so on?

Stuart Russell: Yes, handguns, rifles, machine guns. As you know, you can easily buy 100,000 Kalashnikovs. I think the arms control regime is a really important question, and I know in discussions with the US and others that they often raise this as a reason why it would be difficult to have a treaty. They've refused to ratify other treaties because of concerns about cementenfor. I think you could look at it similar to chemical weapons, where the materials for chemical weapons are readily available, but if you want to

buy large quantities, the chemical manufacturers are required by the national legislation that implements chemical weapons convention, to keep track or they have to know their customer, they have to report any purchases on a grand scale. That cooperation with industry has been quite effective, and I think the treaty, even though it doesn't have very invasive provisions like some other treaties, is generally regarded as a mixed success. I think that you could do similar things with the quadcopter manufacturers.

It'd be quite difficult for a non-state actor to develop the ability to produce millions of quadcopters with the kinds of onboard processing capabilities and cameras and so on that you would you need to make a fully autonomous weapon. If they're only going to do a few hundred, they might as well be remotely piloted, and they're already doing that. The ISIS is using remotely-piloted drones, in some cases Mexican drug dealers are using them to attack each other. So it's really the large scale deployment that you need to try to prevent, and I think with the possible exception of 3D printers, which at the moment are not capable of producing weapons of that kind, they can't produce the electronics, they can't produce the cameras. They can produce the airframe.

Prof Clarke: The fabric.

Stuart Russell: And rotors and possibly the electric motors, but not all the control systems. That might be more possible, but 3D printers themselves, particularly ones capable of producing all the control systems, are not something that a non-state actor's going to be able to get hold of in large quantities. I think there is a feasible issue.

Prof Clarke: Just taking one step back, we were talking about miniature quadcopters here, in your research is there a benevolent use for such devices? At the moment my understanding is 90% of all drones are toys, they're just things to play with. 10% are industrial and differently configured. The idea of tiny drones which make quite a lot of military sense, I'm struggling to think of what the toy versions would actually be. Would I be right in saying that the production of millions of tiny drones could only be for a malign or military purpose? Is that a reasonable assumption? I'm trying to think of whether they could be toys in any meaningful way, that they would therefore be diverted to other use.

Stuart Russell: It's a good question.

Prof Clarke: At the moment is really my question, in what you're aware of at the moment.

Stuart Russell: If you're a toy manufacturer in spirit and you order a couple of million of them from DGI, for one thing they wouldn't come with shaped charge explosives and they wouldn't come with the chips that allow fully autonomous operation, you know, object detection, tracking, application and so on. The toys right now are all remotely piloted, and it might be reasonable

to say they should stay that way. No one has a reason to have a toy that's fully autonomous, let alone one that carries a charge.

Clive Lewis: What size are we talking about, in terms of the micro size drones?

Stuart Russell: The ones you can buy right now are just about 1.1inches in diameter, and that would be more than adequate to carry a lethal weapon.

Prof Clarke: Although they are also very subject to weather conditions, because they're so small. They can't move against a headwind of any sort, can they?

Stuart Russell: They can go about 32 mph.

Prof Clarke: Against a headwind? I mean, they're like birds, aren't they? They get pushed by the wind.

Stuart Russell: Well, if the wind is more than 32mph, yes they wouldn't be able to make progress. Range is an issue, and that will gradually improve with battery technology. Right now they can fly for about twelve minutes, but that's still enough to get you a couple of miles. You would use larger conveyancing drones to carry a couple of thousand, and larger drones have ranges of 50 miles.

Lord MacDonald: You just release them?

Stuart Russell: Yes. I don't want to do the job of the weapons designers for them.

Q5. Prof Clarke: I suppose what we're interested in is how far away we are in time from the realisation of these things. The potential is always very interesting and chilling, as you mentioned. I suppose we're all asking ourselves how close are we to that, in scientific and operational terms?

Stuart Russell: I think it would depend on the level of engineering effort and your ability to just bring together high quality, technical people. I talked to my colleagues who work on autonomous flight quadcopters, and they see this as something that a competent group could produce, and a working system that could then be fielded in large numbers in eighteen months to two years. It's really not a basic research problem.

Prof Clarke: So a competent non-state group is the implication?

Stuart Russell: That, no. I would say it's very unlikely that, for example, ISIS will be able to assemble the competence to do that. I think it's something-, not even something on the scale of the Manhattan Project, but when you look at what they did in about two years, that shows you what can be done when you get the right scientists and engineers behind the project. Right now Lockheed can't hire a good AI machine learning and computer vision person to save their lives, because Google

is paying them $1 million a year. If this was a national priority, and in China and Russia it is a national priority, in the US at the moment they're not on that kind of footing-,

Lord MacDonald: The Chinese have just made it a national priority, haven't they?

> **Stuart Russell:** Yes, so China and Russia has been pushing this. I think Russia's a way behind, but Chinese quadcopters are the best in the world right now, you know, DGI has, I think, 70% of the market.

Clive Lewis: I wanted to come back to the philosophical components of it, but just a question, we seem to be thinking about clunky little machines. I'm just wondering if you know much about, in terms of, the shift into going from micro scales to nano scales, and what the technology is there. Obviously nano technology is something which is a science fiction at the moment really, although there are some manipulations at that level. I just wonder in terms of making the shift from the bigger, clunky machines to the smaller micro machines, which are almost invisible, is that something on the horizon? I'm sorry, I just wanted to, kind of-,

> **Stuart Russell:** Not that I know of for autonomous flight. The smaller you go, the harder it is to generate enough velocity to overcome wind. So going much below 1in, it's not going to be fast enough and it's not going to be able to carry enough power to last.

Clive Lewis: I was thinking an intelligent pathogen, but that's-,

> **Stuart Russell**: Right. Often people ask, 'Why couldn't you just have a mosquito that injects people?' Then you're violating the chemical weapons convention, so those are already banned. It's odd, in fact the chemical weapons convention banned anything that, for example, renders somebody unconscious. Even though you might think it's a pretty humane way to fight a war, that we should just knock everyone out for 24 hours and take over the city, in fact, that's not allowed under this treaty.

Q6. Clive Lewis: On the ethical issues, I was interested in some of the reasons why it has become more acceptable, why we could or would do this. I didn't hear you, you may have come onto it after I left, but one of the reasons screaming at me not to do this, being an avid science fiction follower and watcher and reader, is the technological singularity creating autonomous weapon systems which have no compunction of human beings. I just feel to myself that with machine learning and where we are heading, are we not leading ourselves down to where we're going to be superseded by-, potentially put ourselves in danger? That's the scientific popular end of it. I want to know it's not impossible.

> **Tom Simpson**: Yes, so there are philosophers that speculate around the edge here, and if you've got an infinite threat, like where the world ends when you do this, then the very small possibility renders it irrational to step down that path. I suppose my lay on the side of technology is that I've always found

it quite implausible that we don't end up with that loss of control, which is often listed in science fiction.

Prof Clarke: Implausible? Implausible that we would do that, or implausible that it could happen?

Tom Simpson: That it would happen.

Prof Clarke: That it would happen, yes.

Tom Simpson: Yes. I defer to Professor Russell on the predictions on where we're going, but I know a lot of people in the community do worry about that. I'm always surprised by that. In terms of it becoming more socially acceptable, so the precursor technology that will make this very salient in society is autonomous vehicles. The first autonomous test vehicle driver died in a crash, but there's a plausible defence that he was on the loop, he was able to override the car's autonomy, and it was on his own lookout. The test moment will come where there's a pedestrian walking across a zebra crossing and a self-driving car ploughs in to that person, and obviously that will then be tremendous outrage, salient outrage, 'This autonomous system has killed my child, and why has this happened?' I don't decry the emotion that will arise in that situation, I suppose I just think that the value across the whole of society when we see the number of deaths and car crashes come down dramatically, and the ease of transport and all the efficiency rated benefits in technology promises to bring, we'll be able to recalibrate and say, 'It is terrible for every individual that that happens, but that's part of the cost that we as a society take for technologies that actually, not only make cars better, they make our lives safer in aggregate.'

Autonomous weapon systems are, again, an even sharper version of that, where the emotional salience of being killed by a system's very strong but I think so long as it's plain that the overall casualties' come down, there's responsive administration of these systems, I think that pathway is a perfectly plausible one. So I'm more worried about the unintended consequences of this pathway of development, than the actual act of killing.

Q7. Prof Clarke: Okay, so I can just come in on one quick one, and then we'll go to Lord Macdonald. We've let pass the statement, I think, that Professor Russell made, that the issue of autonomy can mean that no one is responsible. That was my understanding.

Stuart Russell: No. What I meant by lack of attributability was, in the same sense that cybercrime is unattributable, you can't find out who did it. Particularly in a civilian context, since you can target by identity, you can say, 'I want this person. Here's their Facebook page, find that person and kill them.'

Prof Clarke: Sorry, I understand. It's a problem of that attribution, yes. It's not beyond the wit of us to say, wherever, say the United Kingdom. thinks it might use an autonomous system, if we get to that point, that there is a legal requirement that somebody is made responsible for the algorithms that that system operates under on any given day.

Lord MacDonald: There is, in fact, always a chain of legal responsibility. The chain of legal responsibility exists, whether a legal framework has been set up or not. Unless these systems become entirely autonomous, ultimately people control them and they bear legal responsibility.

> **Stuart Russell**: Yes. I find the accountability argument not a convincing reason to ban these weapons. If you define a mission for the weapon, and as a result many more civilians are killed than militarily necessary, then that would be a war crime. Just in the same sense, heat-seeking missiles have a degree of autonomy in choosing a target, but there are scenarios where it is fine to use them, if you're in air-to-air combat, and the only thing in that direction is an aircraft. You can fire the missile in that direction and it will find the target. If you fire it in the general direction of Ealing, it will find a target somewhere in Ealing and that will be a war crime because you're using it in a way where its discriminating abilities-,

Clive Lewis: Tell me if I'm wrong, but doesn't the history of humankind and warfare suggest that we will abuse this technology?

> **Stuart Russell:** Of course, yes absolutely. I had a meeting at the White House where we were talking about these micro weapons and someone from National Security Council said, 'Of course, we would never develop weapons like that.' Well that's not the point. In that case, why don't you ban them? If you don't ban them, maybe your adversaries will develop them.

Prof Clarke: Sorry, we're going to come on in a second phase, to the question of treaty, what we might do and so on. I just want to say, is there anything else we should cover, or that members of the panel want to cover, on the description of what we're talking about, or the philosophical side? I do then want to move on to this question of a UN convention and whether it's an arms control approach or a treaty approach.

Lord MacDonald: We don't have time for the question about strategic arms segregation, say for example, for the United Kingdom.

Prof Clarke: I wouldn't mind getting onto that now. Could we do that? I just want to get all these questions out before we then think about conventions and arms control.

> **Stuart Russell**: If I might mention, I differ slightly to Tom on the question of the acceptability to the general public. So you may recall the picture of the small Syrian boy who drowned on the beach, and the impact that had on European policy. Now imagine that he was running down the beach, being

chased by an autonomous weapon and it blows his head off, and that video is on everyone's television screen. Imagine how much greater the impact of that kind of thing would be. I think it comes back to this predation, the sight of a robot that's chasing the child that's trying to hide or get away, and the robot is relentless, those kind of things have a huge impact on public opinion. I think then you start to get onto the territory of the Martens Clause, where it dictates the public conscience might become very clear.

Prof Clarke: So strategic issue, where's the national interest in this? If we live in a world where these systems are being developed where the Chinese have announced artificial intelligence as a critical national imperative for them, and they're putting billions and billions of dollars into it, and will undoubtedly develop autonomous systems, where's the British national interest in that situation? What do we do? What's the response? Tom?

Tom Simpson: Can I just very quickly come back to-, so the attribute, there's a really important discussion to be had, and there'll be a whole range of views about the acceptability of autonomous systems. Public opinion is pretty fickle and moves on things that are not always-, you know, why that photograph should have the impact it did is a question. I guess the question is the baseline, so it's not somebody drowning that's being chased by a killer autonomous system, it's a question of a person chasing a little child and killing them, versus a machine chasing a child and killing them. So in terms of the strategic national interest, I mean, the CCW process can't impose a ban, because that legal instrument doesn't exist internationally, I'll be corrected by lawyers. What it provides is a convention that nations can voluntarily sign up to, so within their own jurisdiction, they create legislation within their jurisdiction. So the landmine parallel connection for landmines obviously has some non-signatories and some important non-signatories, you know, China and Russia.

So the important question is not 'would we like to live in a world where somebody signs such an instrument?' the question is, 'in the existing world, who will not sign that instrument, and does that expose us to unacceptable levels of risk?' It's for the technologies to really say what the level of asymmetry from a military perspective will be. So landmines, you can wear the risk of not having landmines because they're not actually a very useful system. They bolster you in defence, but they're basically useless in offence, and there's a lot of work involved in picking them up. So we saved ourselves a little bit of that. So the cost, the social benefit of foregoing landmines is very significant, and the military cost is not very significant. What we need is a discussion from the people who are developing the arms and technologies to say, 'Genuinely, what would be the technological asymmetries, the military asymmetries from one adversary possessing these systems and the other not?'

In the event that we don't have assurance that we will not be facing enemies that do possess that technology, there's a compelling interest for us to make

sure we're at least not left behind. We obviously don't want to be accelerating an arms race, but it strikes me there's a very imprudent future-

Q8. Prof Clarke: Sorry, I was going to say, but if we assumed no restraint on the development of autonomous systems and robotics, what would you say would be the UK's strategic position? Would it be one of initial advantage, followed by permanent disadvantage as bigger powers then develop? Is it something like that? If you assume unrestrained growth, where would you say that would leave the United Kingdom, and for that matter the United States?

Tom Simpson: So I'm not confident to give a really very strong judgement on that. As far as I'm aware, we've got a decent but not world-leading AI technology community.

Lord MacDonald: I think people might argue it's a bit better than that, actually. It is certainly as good as many other parts of the world, is that right, or you wouldn't say so?

Stuart Russell: I would say at the moment it's probably second to the United States. China is actually catching up pretty fast.

Clive Lewis: So there is a race already happening?

Stuart Russell: Yes. The race, I think, is primarily because, as Putin correctly pointed out, whoever leads AI will eventually lead the world. Primarily for economic reasons.

Clive Lewis: So in terms of the spin-off, so for example if this is a militaristic arms race, there will be obviously history of warfare (inaudible 47.31) there, you get from a combustion engine upwards you have, you know, commercial spin-offs from the technology that's developed for military applications. Do you mean in that sense? Do you mean in terms of the commercial spin-offs that'll come from developing the military applications?

Stuart Russell: No, at the moment military AI is far behind commercial AI.

Clive Lewis: Can you see it staying that way?

Stuart Russell: Yes, I think. Unless there is a Manhattan Project. Just because primarily I think the quality of people and the level of incentive is so much greater in the commercial sector.

Lord MacDonald: It's certainly true in the US context, is you want to work for Google or Apple, but the question is whether China will - whether the division of labour force and commercial incentives- will be developed in the same way.

Q9. Clive Lewis: Talk us through some of the tactical advantages of an AI versus non-AI combative situation.

Stuart Russell: So actually I want to come back to the question of just the UK national interest in general. So first of all, even if there is a ban or even if the UK decides not to proceed with the development of weapons and other countries do, there is nothing that will prevent the UK from developing anti-robot weapons. Anti-robot weapons are not lethal. So the only caveat to that is if it turns out that the only effective anti-robot weapon is basically an identical device that goes up and has millions of little dog fights with the other devices, that's not the way things are looking at the moment. We're looking at things like EMP, ground base, laser weapons and various other things. My understanding is at the moment there is no effective defence. A ground base laser can deal with an attack by a few tens of weapons, you can easily overwhelm that when you go up into the thousands or tens of thousands. I think there's legitimate argument. I had a meeting with Henry Kissinger and he made this argument, he agrees with my position on the anti-personnel weapons that we absolutely need to avoid, the scalable weapons of mass destruction. He argues, and I think there's some value in this point, that the same considerations don't necessary apply in, for example, air-to-air combat, summary in warfare and so on, where there aren't that many targets, and we're talking about big, very expensive objects. There I would say it's probably true, that if you want, for example, UK to have parity or air superiority and you don't have autonomous fighter jets, then you're not going to have air superiority.

Prof Clarke: Because of the speed of which they can respond and react?

Stuart Russell: Yes, and the physical constraints, the accelerations can be much higher. You can have much smaller because they don't have to carry humans they can be smaller.

Q10. Clive Lewis: Because you're here I'm desperate to ask this question. Answer very quickly if you can, because we mustn't go off on a tangent. How do AI and laws affect our renewal of Trident and the Vanguard-class submarines, just very quickly, in your opinion?

Stuart Russell: I don't see a strong link between these two things. That would be outside my area of expertise.

Prof Clarke: Okay. Even though it'll be in service for 50 years? So it'll be in service in 2060-something, so maritime robotics will not have neutralised it by then? I'm not asking you to predict, I'm just making the point.

Stuart Russell: In terms of a defence ability for submarines, I think we have to assume that the oceans will become completely transparent. That at all times, everyone will know where every submarine is because you can deploy millions of subsurface sensory devices that can assist for years and give you a complete picture. So that notion of hidden submarines lurking around somewhere off the coast, it's going to go-, so in that sense the submarines will

be vulnerable, I don't see how they would really survive. This is not my area of expertise though.

Prof Clarke: No, it just strikes me that-, because AI and robots may be, as they say, the famous dreadnought moment where, you know, in 1906 Britain produced a dreadnought battleship which is the best in the world, and it negated 150 years of naval superiority because in ten years Germans had dreadnought battleships of the same calibre and so on. You are the first in the field with a fabulous technology, which actually neutralises all of the advantages you used to have, but you're driven to have to do it because if you don't, somebody else will develop it. Therefore, that gets you into the issue of, if this is a dreadnought moment with AI, is the most self-interested thing we can do to try to create a treaty of limitations, in our own interest? That's the argument, isn't it?

Stuart Russell: I have to say, this is my view and it's very similar to the biological weapons convention. The US had a big lead, they were spending the equivalent of about $3 billion a year on research and development of biological weapons, and then the biologists and then Henry Kissinger convinced Johnson and then Nixon that the most likely outcome from the commoditisation of biological weapons would be that the US would be attacked, and very hard to defend against. I think there are quite a lot of similarities between that situation and the current situation.

Tom Simpson: So one of the real challenges will then be that the military theorists have been promising transparent battle fields for about twenty years now, but in all my experience, the fog of war is a very present reality even if we may be on the edge of these fundamental technological shifts.

Prof Clarke: I take the point. The fog of war is awfully ubiquitous and you've been in it yourself.

Stuart Russell: I would agree with that, just the oceans themselves. The oceans are becoming transparent.

Tom Simpson: The moment you accept the principle of autonomous systems doing air-to-air combat, the outcome of which decides air superiority, is very compelling. It's very difficult to see the problem there. Then you think, 'Well okay, if that's the case there are clear boundaries, there are these other domains in battle space where we can draw clear boundaries, why shouldn't we have autonomous systems engaging there?' So the issue of where you draw a line becomes a very pertinent one, as to if you accept the principle at all, can you resist it, having full scale?

Prof Clarke: Okay, thank you.

Q11. Lord MacDonald: Can we talk a little bit more about treaties? So Professor Russell, the UN has convened a group of government experts, I think, to consider autonomous weapon systems, and I'm wondering what you think the right outcomes

from that discussion would be. What are the technical definitions and verification regimes we need to develop to enforce a treaty, even supposing states came to together to agree one?

> **Stuart Russell**: It's a good question, and I will be giving evidence on the first day of the next meeting in Geneva. I think the main thing is to cut through the obfuscation about whether it's possible to define autonomy I think that's a relatively straightforward issue from a technical point of view. From this meeting I think we have to have very restricted ambitions. It barely happened at all because first Russia and then Brazil didn't want to pay. It's bizarre, but payments of $15,000, $20,000, or lack of payments, are enough to derail these meetings. They had to do it in November rather than August because it's cheaper to do it when they're all going to be there anyway for the annual review meeting. So that suggests to me that the necessary consensus to move anything forward in the CCW was very tenuous at best. The US, I believe, changed its position. Two and a half years ago I was at Geneva for the 2014 meeting of experts, and there the US was really saying there was no need for any treaty, everything that was an issue was already covered by IHL and everyone should just go away and let the US get on with it. So my views weren't particularly popular with the US delegation.
>
> Now I think the US-, the stated policy letter from Obama was clear that there may well need to be regulation and/or treaty in this area. I think that was one of the steps that made it possible to move forward with the GGE. My sense from talking to other delegations last time is that the UK has a great deal of weight. It's not considered to be as aggressive and jingoistic as the US, but it's considered to have technical expertise and gravitas, and people will respect the position that the UK takes. I think if the UK supports the process of moving forward to establishing definitions, and then the question of verification, once you get the scientists involved in solving the problem-, it's like the advice people always give you about lawyers, never ask a lawyer if you're allowed to do something, because they'll always say no. Ask them how to do it, get them involved in solving the problem, and then they'll figure out a way to make it happen. I think it's the same here, once you get the counties to agree that, in principle, if we can come up with appropriate definitions and suitable verification regime, then a treaty is a good Idea, then they start working on those things so you can initiate that process.

Lord MacDonald: Do you think there are any likely to be any particular definition or verification problems in this space, compared to, say, for example, biological or chemical weapons? Are there things here that would make it trickier?

> **Stuart Russell:** Well, biological weapons doesn't have a verification framework, I believe. That allowed the Soviets to keep going with their programme for a long time.

Prof Clarke: Chemical does.

Stuart Russell: Chemical does. A former member of the OPCW, who is now working on autonomous weapons, explained to me how important the industrial collaboration was for controlling chemical weapons. I'm absolutely confident that the major AI players and people making quadcopters and so on would go along with this, with the possible exception of the Chinese manufacturers.

Prof Clarke: That's quite a big exception, isn't it?

Stuart Russell: Yes.

Lord MacDonald: So you don't foresee any special technical problems in this area, in terms of definitions and verifications?

Stuart Russell: No, I don't think so. The kind of issue that's going to come up is going to be-, so let's say countries continue to develop remotely piloted vehicles, which already have a lot of software onboard to do vehicle health maintenance and help with navigation to deal with weather and keep the flight trim correct, and all the things that military people know about. People would worry that you could easily keep modifying the software on those kinds of weapons, so that they could just operate autonomously. It seems to me that, at the moment, when you have remotely piloted vehicles, because you need a substantial ground crew, you are typically using them with large payloads, and I think they're 25ft long, they carry 500lb of payload, or something like that. That's not going to turn into a weapon of mass destruction, so even if they were operating autonomously, that wouldn't create a threat. That would be a kind of issue that people would raise, how do we verify that something that appears to be a man or remotely piloted vehicle couldn't be converted just by software reprogramming, into something that's autonomous?

Even if we inspected the software in the countries that are willing to make this software available for inspection, how do we know what they get replaced with? So those are the kinds of things that-,

Lord MacDonald: They're quite big problems those, aren't they, in terms of holding people to account?

Stuart Russell: Yes, right. So the treaty engineers, the people who are expert in negotiating these details, getting countries to agree on what can be inspected-, for example, somehow the US managed to convince the Russians to put a US location system on each of the Russian mobile missiles, that sent a signal to the US on a regular basis, just verifying that that missile was in its allowed geographical region. So if you can do that, you can do a lot of things. It turned out that that location detector, by modifying the exact timing of the pulse, was actually communicating the precise location of the missile to the US on a regular basis. So I'm not really an expert in how exactly you can get past these kind of obstacles.

Prof Clarke: The lesson for arms control, isn't it, on verification? Although you can never be sure that you can't stop individual violations, violations have got to be big enough and systemic enough to make a strategic difference, and so you put your faith in the idea that if it's big enough and systemic enough, we'll spot it, even if we won't spot individual violations, necessarily.

> **Stuart Russell**: Yes. I would also say that the risk of that kind of repurposing or reprogramming seems, to me, to be much less than the risk that you would carry yourself, that if you stop basing your defence structure on autonomous weapons, that your weapons will be infiltrated, that people will figure out what the algorithms are and modify the algorithms. So you might think you have a defence posture, but in fact your weapons are working for the other side. That's not true for manned systems.

Q12. Baroness Stern: I don't know if war has been getting less or more in any given period, because it's very hard to tell because what now counts as a war, and how many people do you have to kill in order for it to count as a war, and so on. Well, I just wonder, it would be better, I suppose, if we had less or no war. I just wonder if there's any way with what we've been talking about, and the creation of these systems, could this is any way lead to less war?

> **Tom Simpson**: I'll give it my best shot. So there is a strategic dynamic that could replay in this context which would be akin to that of nuclear. So the logic of mutually-assured destruction has meant that we haven't had nuclear engagements, which is a very good thing at the time. So there is a prospect where the systems become too powerful and we will know that everyone has got them, then conflict becomes irrational, except in circumstances where you know at least one side doesn't have them. So that's a possible future.

Clive Lewis: I was just going to say, it's interesting because there was an essay written by scientists, including Einstein, after the Manhattan Project, I think it was in the early '50s, saying basically that we can't legislate this kind of technology. Ultimately, we have to end wars, if we don't we end ourselves. This does seem like, potentially, from what we've been hearing, potentially another existential threat, another weapon of mass destruction.

Prof Clarke: I wonder if we may be in this binary choice, which is what we face with nuclear. Cyber warfare has been regarded as something which can run out of control, and already we're moving into a deterrent posture with cyber. We say, 'We'll do a bit of defence, a bit of cyber security,' but also in the last three of four years, Britain and America and other western countries have said, 'We have an offensive capacity and we're creating a deterrent logic, so that if we're attacked we can take down your systems as well, and we know that we can do this.' I wonder whether, if in this scope that we're looking at, is it becoming a binary choice as well? Either we have a treaty regime or some combination of a treaty regime, and a deterrent posture. Is that possible? Is that plausible?

Tom Simpson: The cyber case is quite instructive in this because what we haven't managed to do is what Russia and China have been very effective at doing, is creating these cadres of activists who are, in some sense, associated with, not quite taking orders from, but there's a denial relationship, who cause tremendous damage to all sorts of infrastructure. In this instance there is a similar threat of leakage of technologies from central control state organisations to bodies, not working in any controllable way. We're never going to commit ourselves to the model, that structure, internally as a response. So there's a deep security dilemma facing our situation.

Stuart Russell: I think if cyber warfare and cyber crime is the model, then we're all in for a pretty terrible time. At the moment the ordinary person is largely unprotected, and the corporations are unprotected. For many CEOs this is the biggest worry, that they'll wake up and their company will be taken down. I think there's a difference from the nuclear situation. I would also say we survived the nuclear situation by luck, extreme luck. It's almost uncanny how we got through some of these near misses.

Clive Lewis: Sorry, is that right? I'm sure you know more about it than me, but it sounds to me as though mutually-assured destruction ought to work, because then attack in the face of it would be completely irrational.

Prof Clarke: Do you mean, the near accidents that we had during the Cold War, of which there were quite a few.

Stuart Russell: Yes. I mean, the mutually-assured destruction doesn't, I think, apply in the case of these autonomous weapons, because of the scalability issue. You can dial up or down the amount of destruction, and I think it's also true that if nuclear weapons fell into the hands of certain terrorist groups, they might well be used. I think that these weapons are so much easier to proliferate once they're mass manufactured. It's legal, it's not constrained by treaty. I don't feel any sense of reassurance from the history of mutually-assured destruction.

Q13. Lord MacDonald: If you've just got one minute I just wanted to ask Dr Simpson if you, just very briefly, could give your response to what Professor Russell has said about treaties, because I think you're a little more sceptical about them. Just in a minute or two.

Tom Simpson: So I'm not engaged in the details of the process and how these things get worked out, but I suppose one of my thoughts is that there has been a bit of a binary quality at the moment about the debate from the outside, that it's either a complete ban on development of war autonomous response systems, versus throw our hands up in the air and everyone gets on with it as quickly as possible. I suppose the record of a lot of these successful initiatives have been incremental, so find something that you can agree on, implement that, and see whether the norms arise as a result, and then step into something beyond that. So I've been arguing that no first use is a

plausible norm that will be stable internationally, and you might find that if there was a commitment to no first use, certain people will interpret that as, 'Okay, we've got permission to get on and develop them,' but you may see an incremental-, now that the norm of no first use is in the process now, we can agree that there are other areas in which we will strengthen and increase the protections. You may find that a process delivers more, rather than one bang moment.

Prof Clarke: Okay, thank you. Professor Russell or Dr Simpson, are there any other headlines that you want to leave us with, or is there a thought that we haven't fully encapsulated of your analysis, that you'd like to make sure is on the record for us? I always want to give the witnesses a chance to say one last thing, if there is anything. If there is anything, it's not compulsory.

Stuart Russell: No, I think it's been a very good and wide-ranging discussion, and you covered pretty much all the major points.

Tom Simpson: I have nothing further to add.

Prof Clarke: Okay. Well, I have to say thank you very much. It has been a fascinatingly depressing session.

Lord MacDonald: Not all depressing.

Prof Clarke: No, not all. You know that things are changing when old jokes start to seem realistic. There's a famous old joke that says you shouldn't anthropomorphise computers, because actually they don't like us very much. One suspects that with AI we're moving towards that position. Thank you both very much indeed, and we will certainly absorb a lot of this into the report that we write. Thank you.