HOUSE OF LORDS

Artificial Intelligence in Weapons Systems Committee

Corrected oral evidence: Artificial intelligence in weapons systems

Thursday 15 June 2023

3 pm

Watch the meeting

Members present: Lord Lisvane (The Chair); Lord Browne of Ladyton; Lord Clement-Jones; The Lord Bishop of Coventry; Baroness Doocey; Lord Fairfax of Cameron; Lord Hamilton of Epsom; Baroness Hodgson of Abinger; Lord Houghton of Richmond; Lord Mitchell; Lord Triesman.

Evidence Session No. 9 Heard in Public Questions 120 - 132

Witness

I: Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley.

USE OF THE TRANSCRIPT

1. This is a corrected transcript of evidence taken in public and webcast on www.parliamentlive.tv.

2. Any public use of, or reference to, the contents should make clear that neither Members nor witnesses have had the opportunity to correct the record. If in doubt as to the propriety of using the transcript, please contact the Clerk of the Committee.

Examination of Witness

Professor Stuart Russell.

Q120 **The Chair:** Good afternoon, Professor Russell. It is extremely kind of you to join us, and I hope the timing of the session means that it is not too beastly a time for you in a very different time zone. The session that we are having this afternoon is being broadcast and you will receive a transcript so that you can correct errors or, indeed, add to your evidence at any stage. I am aiming to have this session last for about an hour, if that is okay with you.

We are all very well aware of the views that you have expressed publicly about AI and about the hazards, but would you just like to get us going with your view of the UK's policy on AI in weapons systems?

Professor Stuart Russell: Thank you very much for the invitation to speak today. I am sorry that I cannot be there in person. I just want to say a few things before I launch into an answer to the question. Some of my answers today might appear to be critical of UK policy in this area, but, having been involved in the discussions since around 2014, I have always found the UK to be a good-faith actor in the deliberations.

As I said in my Reith lecture, this is not about the general morality of defence research. We would all prefer to have no wars, but, if your taxes are paying someone to die in your defence, it is hardly a moral position to refuse to help protect them. In all the discussions that I have had with members of His Majesty's Armed Forces, I have found them to take a very thoughtful and honourable approach to grappling with these issues, and there seems to be a great reluctance to allow an algorithm to decide to kill human beings, so I am very pleased that these questions are being taken seriously and that we are having this hearing.

The Government's policy, which was stated by Lord Astor in 2013, is that the Ministry of Defence has "no intention of developing systems that operate without human intervention". In the 2021 meeting of the CCW in Geneva, the UK's position was that it opposes "the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifecycle. The use of such weapons could not satisfy fundamental principles of international humanitarian law".

In principle, it sounds marvellous that we really have nothing to discuss here, because the UK has already renounced the use of autonomous weapons, but the devil appears to be in the details of exactly what is meant by "autonomous weapon", and one has to ask, "Why is the United Kingdom opposed to a legally binding instrument that would ban the weapons that the UK has said that it would never use?"

That is a difficult question, and it seems that the UK is equating a weapon that can satisfy IHL with a weapon that does not need to be banned. Of course, a weapon that cannot satisfy IHL is already banned, so the UK's position is that we do not need any new arms control treaties at all. It

also seems to be the case that, if we accept that autonomous weapons systems could be strategically decisive in a conflict, the UK's policy is one of military suicide, since we are insisting that our enemies be allowed to develop and use autonomous weapons against us.

The only way I can make sense of all this is to look into the way the UK is defining autonomous weapons, which seems to be considerably different from the widely accepted definition from the United Nations, the United States and many other participants in the Geneva discussions. The UK definition talks about machines "that are capable of accepting higher-level intent" and relegates most of what the rest of the world calls autonomous weapons to what they call automated weapons, so any weapon that follows an identifiable set of instructions or rules would be automated and not autonomous, and it is only autonomous if its internal operations are mysterious.

That led Michael Fallon, in 2016, to say, in response to a letter from the AI community asking the Government to consider a ban, "Fully autonomous systems do not yet exist and are not likely to do so for many years, if at all", and he said, "It is too soon to ban something we simply cannot define".

The Chair: It rather sounds that, if there is an asymmetry between the United Kingdom definition and definitions that might be employed by other actors around the world, we might be the victims of pedantry. Is that too high a way of expressing it?

Professor Stuart Russell: This is the only way I can understand the UK's statements that it will never develop and deploy autonomous weapons. It is using the words "autonomous weapon" in a way that is a much higher standard than the rest of the world, so it is not denying itself anything except for weapons that may never exist.

The discussions can accommodate different definitions, but if you then say, "We're discussing a ban on autonomous weapons", different sides mean different things, and you could certainly amend the language of the ban to say, "We're banning autonomous weapons that meet conditions A, B, C, D and E", but if you start with a definition of autonomous weapons that is already extremely hard for any weapons system to meet, because it requires human levels of intelligence for example, it is hard to write the definition of a ban because you cannot further qualify what you mean by "autonomous weapon" because you have already defined the empty set with your definition.

The Chair: Do you see any signs of convergence in the international defence and scientific community towards a definition that would be workable in the way it is handled internationally?

Professor Stuart Russell: Yes. The United Nations' and the United States' definitions are pretty similar. They talk about weapons that can locate, select and engage human targets without human intervention.

Q121 **Baroness Hodgson of Abinger:** Professor, you are tiptoeing into this

area already. My question is around standards and international humanitarian law. What international standards of practice on the development, testing and use of AI-enabled and autonomous weapons systems are needed to ensure compliance with IHL? How are these standards going to be developed and who will develop them? How would they be future-proofed?

Professor Stuart Russell: This is an interesting question, and there has not been very much discussion of it that I am aware of. If you look at the United States' recent political declaration that was announced in February on the responsible use of autonomous weapons, almost all the requirements, except for the one on nuclear weapons, which we could certainly discuss later, apply equally well to the manufacture and sale of sandwiches. In other words, they are very general. They talk about paying due care and attention to quality and accuracy and making sure that they do not go mouldy while they are in the fridge and so on.

The particular difficulties that I see in testing for compliance with IHL would be proportionality and necessity. The discrimination requirement is similar to the kinds of things that we already test for with AI systems, and the particular issues there would have to do with robustness to adversarial objects, an adversarial object being one that is specifically designed to fool a machine classification algorithm.

It has been demonstrated repeatedly that with an algorithm that is designed for example to recognise thousands of classes of objects, such as an ostrich, you can take any other object such as a school bus or a horse and make invisible modifications to the image, and the algorithm is convinced that it is an ostrich, or vice versa. We can do this not just with photographs but with real physical objects. You can take a 60-mile-an-hour sign and convince the algorithm that it is a stop sign or a yield sign, making more or less invisible changes to the physical object. People have shown that you can apply fairly invisible surface markings to a small plastic turtle and convince the algorithm that it is a rifle.

There are difficulties in testing for discrimination, but proportionality and necessity are things that are so context-specific and dependent on aspects of the overall military situation that it would be very difficult not only to design an AI system that could make that judgment reliably, but to develop any of kind of testing in the lab for those conditions.

I am not sure how you would design situations that are fully representative of the kinds of situations that could occur in the field, where there are difficult judgments to make, but one thing that seems obvious is that you would want to have facilities for very large-scale and realistic simulation with what we call red teaming—namely, with people whose job it is to put the AI system into a difficult situation in order to test whether it can judge these properties successfully.

As for who would do this, the written questions that I was sent mentioned the *Tallinn Manual*, which was designed to address questions of applicability of IHL in cyberspace. It was written by an independent group

of academics, so it does not have official status, but it was commissioned by NATO, as far as I can tell, and is used across NATO. I doubt that the North Koreans have decided to abide by the conclusions of that report.

If we are to use autonomous weapons systems, it would make sense to go through a similar exercise of developing common standards. It would make no sense for every member of NATO to develop its own standards. You would need to have AI experts as well as people with a great deal of experience on the ground with actual warfare and how human beings make these kinds of decisions.

I believe that that was done with the *Tallinn Manual*. They had not only computer scientists, cybersecurity experts and legal experts, but people who had operated cyberwarfare centres, so they knew what was happening.

Q122 **Lord Hamilton of Epsom:** Professor Russell, in the earlier deliberations of this committee, I raised the question of the drone attack on a wedding in Afghanistan and why there were no prosecutions. The reason was relatively straightforward—that the barrier that had to be cleared in terms of the charge was that this was a war crime. I reckon that was one reason why people were prepared to sign up to this particular treaty, because the barrier was so high that it was very unlikely that anybody was ever going to be convicted under it.

Do you see this changing when it comes to changing international humanitarian law vis-à-vis AI, or will we have the same barrier of needing to prove that it is a war crime before any convictions or any charges are brought against the operators?

Professor Stuart Russell: That is a very good question. Compliance with IHL is having an effect on the various parties as they consider their development of autonomous weapons. Most people in the AI community are surprised by the slow rate of progress on and introduction of autonomous weapons. Part of that is due to caution. In my meetings with policymakers in the US, for example, people are concerned about violating IHL. The most likely violation that they are concerned about, since the warning by Christof Heyns in his report in 2013, has been discrimination: that a weapon would be launched and would make a mistake similar to the one you describe.

The difference is that one would assume that attacking the wedding was not a deliberate act and therefore that the human commander who made the decision to authorise the strike made a mistake. However, there was no assumption that that human commander was inherently incapable of discriminating, whereas the warning is very clear that these weapons are not yet capable of discriminating. If you were to launch such a weapon and it made a mistake, you would be in a different legal situation, so there is caution in moving ahead and deploying these weapons.

I noticed that Turkey, when it released the Kargu back in 2017, was making all kinds of claims about its ability for autonomous hits on human targets and has since been back-peddling furiously. When I talked about

this in Geneva at the CCW meeting in May, the Turkish ambassador had come with a prepared statement and accused the international community of perpetrating a slander against the Turkish state, lying about the Kargu autonomous weapon, and claiming that it was never autonomous and had never been claimed to be autonomous, and so on. Fortunately, I had made a copy, using the internet archive, of the original adverts for the Kargu weapon, which clearly state that it is capable of autonomous hits on human targets using face recognition and so on, and I am waiting for the Turkish ambassador's reply.

The Chair: It rather sounds as though you took the view that the ambassador was not autonomous.

Professor Stuart Russell: No, possibly not.

Lord Hamilton of Epsom: Professor Russell, can I come back for clarification? You see an attack on a wedding not being classified as a war crime, but you see an AI attack being classified as a war crime, or are you saying that there would be a lower threshold than a war crime for AI?

Professor Stuart Russell: If the weapon is known to be unreliable and incapable of proper levels of discrimination, the decision to launch it would be a deliberate decision to launch a weapon that was indiscriminate. That is different from the situation that occurred in the wedding, where presumably the attack was made in good faith, so to speak, even though it was a mistake.

Lord Hamilton of Epsom: You see western defence agencies launching AI systems that they know to be unreliable.

Professor Stuart Russell: I doubt it. That is the reason for the slow movement in this area. The Manhattan project took about two years from a point where even the physics was not well understood to the point where the weapon was deployed. We are not seeing progress at anything like that speed, despite the fact that the problem here is considerably easier than the one faced by designers of self-driving cars, which have to be far more reliable than a weapon.

Q123 **Lord Browne of Ladyton:** Professor, I want to ask you some questions about a phrase that is often attributed to you—the "sole ownership fallacy" of autonomous weapons systems. I would invite you, in the first instance, to just expand on what you mean by that. What risks are posed by such a fallacy? How widely is it held?

I am really interested in how concerned we should be about the possibility of AWS being used by non-state actors and, in particular, how far national security capabilities should be outsourced to private corporations. We have an example at the moment in Ukraine, where Elon Musk has been said to have his own defence policy because he controls access to Starlink by the Ukrainians, much to their frustration, because he makes a judgment about how much potential escalation he can live with. I would like to hear what you have to say about that, since you have given the whole discussion this phrase, which is very interesting.

Professor Stuart Russell: The meaning of the phrase refers to the assumption, which underlies a lot of the discussion, that only one's own armed forces, and perhaps those of one's allies, would be the owners of this new class of weapons. That was the case for quite some time with Predator drones and related technologies. Clearly, it neglects the possibility that one's opponents would have these weapons and, as you mentioned, that non-state actors would also fire these weapons.

We can see this mistake appearing. Even last week, there was a newspaper article about the development of some particular category of autonomous weapons, and there is this constant reference to the fact that these weapons keep our soldiers out of harm's way. That is simply not true. They keep our soldiers out of harm's way only if we are the only people who have those weapons. If the other side has the weapons, our soldiers are exactly in the harm's way that we hope to put their soldiers in.

In fact, what we are seeing in Ukraine, with the use of human-piloted drones, is that they can increase the accuracy of targeting in two ways. One is quadcopters that simply drop hand grenades directly on to troops who are in trenches. The other is using drones as artillery spotters, which allows the artillery to very quickly zero in on the target. It seems likely that this is leading to higher casualty rates among soldiers than would otherwise be the case, and it is making trenches less and less viable as a means of protecting one's troops. I have read stories about the need now to go underground so that you cannot be detected from above by drones, and by satellites.

In terms of whether non-state actors are likely to get hold of these, the main concern of the AI community—it has been since around 2015—is that, because autonomous weapons do not need human intervention, by definition, one person or a small group of people can launch as many weapons as they can afford. If these were small grenade-carrying quadcopters or other kinds of kamikaze devices, you could be launching these weapons in the tens of thousands, or conceivably even in the millions. They would have the same lethal effect as a hydrogen bomb, yet they would proliferate in the same way that small arms proliferate.

We know that there are in the order of 100 million AK-47s in non-government hands at the moment, so we would expect proliferation on that scale. These weapons would be very cheap. I would imagine that they could cost as little as a landmine; a landmine can currently be obtained for about £6. That would be the situation that we would face, and not only that: as weapons of mass destruction, they would be very much more useful than nuclear weapons, because you could select your targets and, rather than killing everybody, you could select on the basis of ethnicity, gender or age, and you do not leave behind a huge radioactive wasteland.

This is the same argument that the biologists used when they wrote to President Johnson, explaining that if the United States research programme to develop biological weapons succeeded, it would lead to

weapons of mass destruction proliferating that would be cheap, easy to create and very likely to be used against the United States. By pushing ahead with that class of weapons, the United States was reducing its own security and tipping the balance of power away from the major powers and towards rogue states and non-state actors. We are in the same situation with lethal autonomous weapons.

This is the principal argument that the AI community has been making for why we should not move ahead, particularly with the anti-personnel weapons. There are different arguments for the larger weapons—the submarines, fighter aircraft and so on. I speak for many of the leaders of the AI community, because they have all signed open letters stating this argument. We think there is no good case for lethal anti-personnel weapons.

Q124 **Lord Browne of Ladyton:** If I might revert for a moment to my former job and ask you a leading question, is the reason for you, as a policymaker, raising this issue in the way you do and making it so accessible, to ensure that when we consider what are now called the guardrails for the development of this technology, particularly in weapons, we keep at the forefront of our mind the reality that there will be a proliferation of this capability well beyond nation states and others into an environment in which we do not normally have arms control in any form?

Professor Stuart Russell: There are two kinds of proliferation. The proliferation of software capable of directing such a weapon is inevitable and cannot easily be prevented, so one might ask, "What's the point of banning it if it's going to proliferate anyway?" Trying to apply the regulations to software and trying to contain the spread of software is probably futile, but the ban would apply to the physical weapons platform, not to the development of software that can make these types of decisions.

If there were no ban, then, as with landmines prior to the landmine treaty, the weapons would be manufactured in very large numbers and would be obtainable in the international arms markets, which vary enormously in their propriety. If there were a ban, it would be very difficult to have large-scale manufacture of these weapons without detection. I do not think that Isis would have been able to set up a factory producing millions of these weapons. It could certainly have bought civilian devices and converted them to carry weapons, which it did with civilian, human-piloted drones, but in a conversion process at small scale there would be no particular advantage in using autonomy in those cases. At small scale, you could pilot the weapons with human pilots, which, at least for the foreseeable future, will be more effective in any case.

It is really the mass production that we are trying to prevent, and there are reasonable ways of developing the kinds of regime that we have for the Chemical Weapons Convention to prevent diversion of civilian

precursor products into the large-scale manufacture of autonomous weapons.

Q125 **The Chair:** If you are going to have regulation of any sort, if you have a treaty you need to enforce it. That means that you have, for want of a better phrase, weapons inspectors. In many spheres, the inspection of weapons may be politically difficult but, technically, it is fairly straightforward. You have described being able to spot the huge production, perhaps, of drones. That is the hardware expression of there being an AWS in the game, but surely the most difficult thing is to assess the degree of autonomy of a system, which may well be just lines and lines of code. Can you just give us a feel for how one would approach that sort of technical challenge? Is it completely unrealistic?

Professor Stuart Russell: That is an excellent question. The issue is one of distinguishing between whether a factory is producing autonomous weapons or weapons intended for human piloting. It would be a good idea for a treaty also to impose constraints on the physical design of human-piloted weapons. One such constraint would be that there be an air gap, as they say—a physical separation—between any onboard computing in a human-piloted weapon and the firing circuits that can drop the hand grenade or launch the missile or whatever. That would mean that it would be impossible or close to impossible to operate that as an autonomous weapon simply by changing the software.

I agree with the premise of examining the software and trying to decide, first, whether it will run autonomously, and preventing it from being instantly updated once the inspectors are gone, and we say, "Phew. Press the button and reprogramme all the devices to be fully autonomous". It would be really difficult to prevent that, but if you put design constraints on human-piloted weapons, you would also presumably require an appropriate degree of correspondence between the number of human-piloted weapons and the number of human-piloting stations. That would tend to prevent confusion between whether a facility is producing human-piloted or autonomous weapons.

Q126 **Lord Triesman:** My question, really, flows from what you just said and from Lord Browne's supplementary question. I understand the point that the proliferation of software may be inevitable, as you put it, but is it conceivable that by lifting large numbers of lines of code out of AI resourcing systems you could end up with a very much larger number of non-state actors having access the algorithm part of it? Are we, in effect, amplifying the consequences and exacerbating the problem?

Professor Stuart Russell: If I might ask for clarification, what are we doing that is potentially amplifying the problem?

Lord Triesman: I am told now that if I use certain sites I can get lines and lines of code that would enable me to build algorithms to do things that I would not have been able to do before they all became available in that way. Although I understand that it may be very difficult to prevent the development of software, has the development of software become

so easy that malign non-state operators could, in very large numbers, now do this using artificial intelligent sources?

Professor Stuart Russell: Yes, software capabilities are available in open-source form and are in the public domain. If one had the physical weapons platform, it would make it quite easy to create the weapons-control software. That has been true for quite a few years.

The technical difficulty is not that great for doing this and, as I said, it is much easier than building a self-driving car, because a self-driving car has to operate with eight 9s of reliability, whereas a weapon would typically only have one 9 of reliability, if that. The weapon does not have to have a level of common sense to know what to do when there is a policeman and a flock of sheep on the road at the same time. It is really a question of whether non-state actors can acquire or produce the physical platforms in large enough volume to present a real threat.

Q127 **The Lord Bishop of Coventry:** Professor, thank you very much indeed. It is a fascinating session. My question was touched on: the role of private industry. What do you see as the role for private industry, particularly civilian AI developers, in the debate on AWS?

I wonder whether I might just make that a little more specific on the back of some interesting written evidence that we have had from Microsoft. Just by way of example, that written evidence tells us of a recently published blueprint for governing AI deployment, outlining the need for effective safety breaks.

The Chair: I think you are referring to evidence that we have had in confidence. Keep going, but please paraphrase the point you are after.

The Lord Bishop of Coventry: I will have to be very careful in that case. We have had a lot of comment from industry, which you have alluded to, about the need for regulation and a readiness to work with government for it. I was quoting an example of a firm saying, really, "We are offering some ideas here, but we need government to pick it up and to work with us on those ideas, so that they then become implemented across the industry".

I apologise for revealing something about some evidence put in confidence, but I assume that it is well known that a particular firm is trying to do this, and I imagine that it is not the only one. How reliable is that sort of work, which is clearly coming from a commercial company? Can that be trusted and worked with? This whole interface between the private and the public has been quite a theme.

There is a wider question that we want to ask about the involvement of various actors in the debate on AWS, but I am particularly interested in the point about reasonable and reliable engagement with industry in developing the technical needs to make weapons systems safer.

Professor Stuart Russell: This is a very interesting question. It is true that the private sector has a lot of experience now in ensuring that AI

systems behave themselves in civilian applications. We are seeing at the moment with GPT-4 and other large language models that, despite the best efforts of the manufacturers, the systems continue to misbehave. It is quite likely that the technologies underlying large language models are going to turn out to be impossible to control properly, because they are created essentially by an impenetrable training process, and there is no way to put in any guidelines or guardrails on their behaviour other than by spanking them lots of times and hoping that they learn their lesson.

For autonomous weapons, if we take the simple example of discrimination—learning to tell the difference between civilians and combatants, and legitimate and illegitimate targets in warfare—industry has experience in developing large training sets. They have learned the hard way that adversarial examples can make fools out of their algorithms very easily, and one would have to expect that in warfare there would be adversarial examples where tanks would be printed. If I knew the recognition algorithm that a given weapons system is using to detect tanks, I can print on the tank a pattern that would cause the algorithm to recognise the tank as an ostrich or a penguin, or anything else that we choose. There would be extreme value in obtaining copies of the recognition code that the weapons are using in order to defend your own targets and to create decoys. You could then take perfectly innocent objects, like lampposts, and print patterns on them that would convince the weapon that this is a tank and should therefore be attacked.

Even with all the experience that private industry has in building and testing robust systems, it would be very hard to anticipate the kinds of adversarial attacks that would emerge in warfare. There are other concerns that I have heard raised about the possibility of AI systems being biased, and we certainly see that in civilian contexts, but the decisions that are made by autonomous weapons are much blunter, shall we say, than the decision whether to interview a candidate for a job based on their résumé and so on.

Most of those considerations are not particularly relevant, but the style and ethos of work for most companies in the AI sphere is indeed, as the written question says, "Move fast and break stuff", and things are put into the field. Another colleague of mine described it as a bunch of dudes chugging Red Bull working late into the night and producing a piece of software and, the next day, it is released to billions of people.

That style of work and development is very different from what one would see in the defence industry, which, for good reasons, is slow, bureaucratic and careful, and does lots of testing before releasing a new weapons system. We just need a different culture. In fact, IBM, for that reason, had an entirely separate division that carried out its contracts with the defence department in the US for many years, because the required culture was so different.

Q128 **The Lord Bishop of Coventry:** On the back of what you were saying about the "move fast, break stuff" culture, do you have any views on an interesting proposal that was put to us by Professor Gopal Ramchurn—I

think I am allowed to refer to this—from the University of Southampton, which is that people from different disciplines and who have different ways of operating than the ways you have just described should be involved in AI development to ensure that developers adhere to certain principles? Is there a role for this wider involvement, especially where clearly the technical knowledge would not be there but other virtues and wisdom would be brought to the table?

Professor Stuart Russell: For civilian applications, it is beginning to be widely accepted that one needs to have people with a social science background, because the issue is not just, "Does the algorithm work?" but what happens when the algorithm is placed in a particular social context. In fact, that can even have technical consequences; if you put the algorithm into a particular social context, it can affect the social context in such a way that the training data, whereby the algorithm is continuing to adapt to its situation, is changed by the presence of the algorithm.

This is something that lenders have understood for a long time. They call it adverse selection. When you create a lending product, you do a lot of testing in advance to make sure that it is going to be profitable, but when you put the product on the market you find that the people who apply for the products are a very different population than you had in your training data, so you end up losing a lot of money.

The same thing happens with insurance. The availability of an insurance product changes the behaviour of the people who acquire the insurance policy. There is increasing understanding of that for AI systems that are placed in the field. Another example is algorithms for doing triage in hospital emergency rooms. You have to look at how it gets used and what effect it has on practices and so on, which is not something that computer scientists are used to.

Computer scientists are not taught about the different cultures of the people who will use the algorithm or the differential effects on those cultures and so on. There are many reasons why, when you deploy an AI system in a civilian context, you need to have social scientists, ethicists and so on, appropriate to the nature of the system and the context that it is going to be put in.

For the defence industry and weapons systems, as I mentioned earlier, there is a lot less nuance, but much more care and attention still needs to be paid to potential negative consequences, which in some ways are similar to what you need for a diagnostic algorithm in cancer medicine, for example, where a mistake can cost somebody their life. You would need people who have much more experience in high-stakes engineering, which might already be present in the defence industry.

The people who develop fighter aircraft or the weapons systems that belong on them are very used to the idea that this is a high-stakes application, and there needs to be all kinds of safety testing and fault tree development. That engineering mindset needs to be brought to bear. I

am embarrassed to say that in computer science, certainly in the United State—less so in Britain—even the idea of correctness, which you might think is the most basic property of an algorithm that one should care about, is not taught in a systematic way to computer scientists, so you often have to learn it in practice.

Q129 **Lord Clement-Jones:** Hello, Professor. I want to follow up on industry involvement in institutions like the GGE, because it is fairly notable that there is not a great deal of involvement—it may well be your experience—by the industry players in the GGE. I wondered whether you would welcome more involvement, and comment again on the possibility of improving the culture of the industry, so to speak. You talked about the "move fast and break things" culture, but you also said that they had learned the hard way. I do not know whether you are referring to the Silicon Valley mainstream tech companies as a result of things like Project Maven and so on, or whether you are looking more widely at defence technology companies specifically.

What faith do you have in the ability of those companies to change their culture, maybe by reference to individual engineer ethics or whatever it may be? Should they be much more engaged in the debate than they are now?

Professor Stuart Russell: As far as I know, it is correct to state that the CCW does not countenance the participation of corporations in the discussions, so they allow for other international organisations and civil society organisations to participate, although Russia has recently been refusing to allow them to speak, which is a bit annoying.

Lord Clement-Jones: Perhaps I should rephrase that—inputs rather than representation.

Professor Stuart Russell: Input from corporations with respect to their technical knowledge and experience would always be welcome, and that would come from defence manufacturers as well as from software companies. With a few exceptions, the standards of quality in software are not as high as one would want for developing a weapons system. Errors are constantly being introduced and later corrected by software updates that happen on an almost daily basis.

You mentioned Project Maven. There is also in the software industry, certainly in California, a great reluctance to work on these kinds of projects. As I mentioned at the beginning, it is not a generic reluctance to work on defence-related projects. It is a specific reluctance to work on algorithms that can decide to kill people. There is a widely accepted view that that is morally unacceptable.

If I was a national Government deciding whether to proceed with the development of autonomous weapons, the biggest practical problem—I see this in the US specifically—is that the most talented AI researchers will not work on autonomous weapons. Civilian sector salaries are far higher than those in the defence sector. Just the nature of work and the procedures for doing it and so on are alien to most computer scientists,

so it has been very difficult for the defence industry to be at the cutting edge of these technologies. That is another reason why progress has been very slow.

In the geopolitical context, people point out that China is not subject to these kinds of problems and restrictions. That is a question that western countries will have to face if we do get into an arms race in this area.

Q130 **Lord Fairfax of Cameron:** Thank you, Professor Russell. This is really interesting. A witness in an earlier session told us that AI is currently overhyped, that its development is linear, and that it is just applied statistics. Do you agree?

Professor Stuart Russell: Yes, to some extent. It is overhyped, given the hyperbolic language that one sees, but the development is not linear, whatever that means. Clearly, in the last decade or so, there have been huge qualitative advances. Initially, in the early part of the last decade, we saw dramatic advances in object and speech recognition and machine translation. Problems went from being unsolved to solved. Also, the legged locomotion of robots was essentially unsolved and is now solved.

The large language models were completely unexpected. Ten years ago, they were a niche product that was used to clean up the output of machine translation and to help with text completion when you are typing on your cell phone. No one expected that a large language model could convince someone that it was conscious and could spend 20 minutes trying to convince someone to leave their spouse and marry the machine and so on.

The hype is restricted mostly to those who say that these systems are already general purpose intelligence or artificial general intelligence. If you listen to Sam Altman's evidence in the Senate and some of the other speeches that he has given recently, he has been very clear that these systems are not AGI, that they display a number of weaknesses and that there are still big open problems that have to be solved before we reach the level of AGI.

There are those who are claiming, perhaps to increase the valuations of their start-up companies, that this technology is AGI, which it most certainly is not, but there is no question that the capabilities that the systems exhibit are massively beyond anything I would have expected five years ago.

Q131 **Lord Hamilton of Epsom:** Opinion polls show that there is widespread opposition to the use of automated weapons systems, which appears at odds with the UK Government's openness to the development and employment of AWS. How can we facilitate public engagement and debate on the issue of AWS?

I would just add that there has been an awful of comment about the civilian use of AI, with Elon Musk saying that it should be paused for six months and other people saying that it is the road to Armageddon, which clearly is not helping either.

Professor Stuart Russell: My sense is that the public understanding of AI is a bit confused by a lack of understanding in the media, but the public positions, as evidenced by the polls, are reasonable. A majority believes that the future development of AI presents a risk to human civilisation, which I think is correct, for the simple point that it is hard to maintain power for ever over entities that are more powerful than us.

On the issue of weapons, it is improving a bit, in the sense that the media tend not to talk about Terminators anymore. They have gradually stopped putting a picture of a Terminator on every article about autonomous weapons. The effect of putting a Terminator on the article was that people were led to believe that this was science fiction and something far off that we did not really need to worry about. Now, the story is often about an actual autonomous weapon that is being fielded.

The public understand that there are good reasons why we do not sell nuclear weapons in Tesco's. It is not some highfalutin ethical concern. It is just plain common sense. That is the core of the argument for why we should not proceed with autonomous weapons systems: they would be cheap and easily proliferated weapons of mass destruction. It is not an ethical issue; it is a common-sense issue. That is the core argument that the AI community has been making. If the public understand that, that is fine.

There are more nuanced arguments about other categories of weapons. The autonomous submarine is not going to be used as a weapon of mass destruction, and there are few civilians underwater. That is a whole different discussion about stability in cybersecurity and accidental escalation, and various other considerations.

I am not sure what the Government can do to improve public understanding. In the US, the Congressional Research Service publishes explainers, partly intended for members of Congress, so that they can quickly get up to speed on issues. Until last year, the Congressional Research Service explainer on autonomous weapons did not mention the WMD argument. It listed the pros and the cons and did not include the main con argument at all. That has been updated, so now it does, which is good.

The Chair: It sounds to me a little as though you are encouraging this committee to play a role in enhancing public engagement and debate, which is really one of the things that we are here for.

Professor Stuart Russell: Yes.

Q132 **Baroness Doocey:** A lot of the people who I speak to are really worried about AI in general. They see it as almost the end of the world. Do you believe that the positive benefits of AI outweigh all the negatives that we have been discussing today, or do you believe the opposite to be true?

Professor Stuart Russell: It is a difficult question to answer, because AI is not one set thing over which we have no control. We could decide to pursue AI in such a way that we lose control over our own civilisation, in

which case the negatives would outweigh all possible imaginable benefits that might come from it.

When both sides possess the weapons, the military advantages are cancelled out. There are arguments saying that the weapons could be more precise and involve less collateral damage, but that remains to be seen. As we have seen with human-piloted drones, there is certainly collateral damage occurring, and to a much greater extent than was previously understood.

My general view of AI is that people are right to be concerned about its misuses in the marketplace, its risks to jobs and the overall risks to our control over our own civilisation. People have pointed to the complete failure to anticipate and control the negative effects of social media, and have said, "Let's try not to make the same mistake again with more powerful forms of AI".

The Chair: Professor, we will have to leave it there, because the clock is against us. Thank you very much. You have been extremely generous with your time, and you have posed a series of rather sceptical questions, which we will certainly be pursuing as we take our inquiry forward. Once again, thank you very much indeed. It has been a pleasure to have you with us.