HOUSE OF LORDS

Artificial Intelligence in Weapons Systems Committee

Corrected oral evidence: Artificial intelligence in weapons systems

Thursday 29 June 2023

11.05 am

Watch the meeting

Members present: Lord Lisvane (The Chair); Lord Browne of Ladyton; The Lord Bishop of Coventry; Baroness Doocey; Lord Fairfax of Cameron; Lord Grocott; Lord Hamilton of Epsom; Lord Houghton of Richmond; Lord Mitchell; Lord Sarfraz; Lord Triesman.

Evidence Session No. 12 Heard in Public Questions 156 - 164

Witnesses

I: Laura Nolan, SRE and Principal Engineer, Stanza Systems; Taniel Yusef, Visiting Researcher, Centre for the Study of Existential Risk; Professor Christian Enemark, Professor of International Relations, University of Southampton.

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 witnesses

Laura Nolan, Taniel Yusef and Professor Christian Enemark.

Q156 **The Chair:** Good morning, all three of you. Thank you so much for taking the time to be with us this morning. You probably know the form for this sort of encounter. It is being broadcast and a full transcript is taken. You will see the transcript in draft to be able to correct errors of fact. Of course, you can come back to us after this oral evidence session if there are points that you want to make to us in writing. We will go for about an hour, if that suits you. Could you briefly introduce yourselves?

Laura Nolan: I have come over from Ireland. I am a software engineer and have been coding for—gosh, a very long time now—two and a half decades. I have degrees in computer science and artificial intelligence. I have also studied, at master's degree level, ethics and strategic studies. I am currently studying safety science and human factors. I have a broad interdisciplinary view of a bunch of things, although I am not a professor.

I am a professional software engineer and have built a lot of systems that I consider to be autonomous, although not weapons systems. My interest in this area started about five years ago when I was working at Google and was asked to participate in Project Maven, which I assume is well known to this committee. Since then, I have taken an interest in the area and have worked a lot with the Campaign to Stop Killer Robots. That is probably all you need to know about me.

Professor Christian Enemark: Good morning. I am a professor of international relations at the University of Southampton. For the last five years, I have been supported by the European Research Council through its Horizon 2020 research and innovation programme to run a major project on the emergent ethics of drone violence. That has included some consideration of AI technology in those weapons platforms. I am also one of 30 co-investigators working on the Trustworthy Autonomous Systems Hub, which has its headquarters at the University of Southampton and is funded by UK Research and Innovation. I have about 20 years' experience researching and teaching arms control and military ethics at universities in Australia and the UK. Thank you for inviting me to join your session.

Taniel Yusef: Good morning, everyone. It might be interesting to know that my first LLM was in International Economic Law, with International Financial Law and Human Rights Law, when I studied global trade, WTO dispute resolutions and such. In parallel, I undertook associate LLM study in IHL, LOAC and global trade, which may help us when we discuss industry. I have also been a ground researcher in conflict zones for a number of years, most recently in Ukraine, including Bakhmut until Christmas.

Until a month ago, I spent seven years as UK International Representative on Trade, Economics and Disarmament Affairs for Women's International League for Peace and Freedom, the oldest

women's peace organisation in history. It is now 108 years old and has won two Nobel prizes. I am currently a visiting researcher at the Centre for the Study of Existential Risk at the University of Cambridge. I work on AI, unpredictability, outer space threats, emerging economies and various global existential threats. I work particularly on feminist issues. I am the technology developers' co-ordinator for the UK Campaign to Stop Killer Robots. If you include the associate LLM study, my fourth or fifth masters depending on your view, is current undertaking of an MSc in artificial intelligence.

Q157 **The Chair:** Thank you. We engaged with the Centre for the Study of Existential Risk during our visit to Cambridge a few days ago, so it is particularly good to have you with us. Can I start by asking you all about the UK's policy on AWS? Professor, you are on record as saying that in the UK, we ought to move closer to the US and ICRC definitions.

Professor Christian Enemark: When I look at recent documentation, my reflection is that the UK Government have, or at least would rather have, not a policy on AWS but a policy on AI-enabled weapons systems. That seems to be the term they are apparently more comfortable with, and they seem to feel a palpable, discernible awkwardness with the term "autonomous weapons systems" or AWS. They appear to be a bit of an international outlier in coming up with a definition of that term, which is critically important if it is to form the basis for guiding choices and behaviour.

That said, looking at the entirety of UK policy, I have the impression that the UK Government have made a good start. The policy contains lots of sensible ideas, but there is certainly room for improvement. I am very interested in the concept that the UK Government put out, which is "context-appropriate human involvement in weapons which identify, select and attack targets". It is possible that that concept could get some traction and it could be helpful to work through what it means. That work has yet to be done but at least the UK Government are being constructive. Having been visibly uncomfortable with the expression "autonomous weapons systems", they at least offer this other term and we can see how to work with that. It is also heartening to see the UK Government state clearly their view that human responsibility and accountability for decisions on the use of weapons systems cannot be transferred to machines. That is an excellent starting point, recognising that machines cannot be recipients or bearers of responsibility and accountability.

The other thing I would reflect on in the UK's policy on AWS, if we want to call it a policy, is that the UK Government are more than just the Ministry of Defence, and weapons can be used outside armed conflict, yet the conversation has been primarily directed towards armed conflict and the law that governs it, which is international humanitarian law. It need not be so restricted. Arguably, it ought to be expanded to include the use of violence by the state—for example, for law enforcement purposes and in policing functions. We need to think about what the implications of AI incorporation might be in that context. Once we get out of the context of

armed conflict, we are not restricted to talking about humanitarian law; we are open to be inspired and guided by international human rights law as well.

Laura Nolan: First, artificial intelligence is not a particularly well-defined term, so I echo Professor Enemark's comment that it might not be the best place to anchor your policy. The definition of AI has shifted hugely, even in the couple of decades of my relatively short career. Things that were considered artificial intelligence when I studied at undergrad level are now considered routine. The boundaries of what AI is shift. There is no well-defined stable definition of that term. We want policy to have longevity. We do not want to have to update our policy every 10 or 20 years. For that reason, this may be ill-considered.

I underline the importance of autonomy and why we talk about autonomous weapons and not AI weapons in the discourse of the UN. First, autonomy does not necessarily imply artificial intelligence. It implies that there is a decision to use force that is taken by a machine—leaving aside the big philosophical question of whether a machine can make a decision.

The Chair: For example, an anti-personnel mine would fall into that non-AI autonomy.

Laura Nolan: Absolutely. Even something like the Harpy is not really using artificial intelligence. You are probably all aware of what a Harpy drone is. We should not conflate the two things. Autonomy brings its own challenges, separate from artificial intelligence. Once autonomy is happening, you have brought another type of actor into the system. Human beings behave in ways that are typically sensitive to the context in which we operate. You and I, sitting here having a conversation, do not need to fully script it out and decide in advance everything we do. We can be adaptive to each other, our environment and context. That is not true of even the most advanced artificial intelligence systems at the moment. You have to script out what they should do in what context. The machine-learning components are typically about sensing the environment and a target profile. The decision is not context-appropriate, so even very simple autonomy brings challenges.

I will give an example that I have spoken about before, which is something I heard Nancy Leveson talk about. She is a well-renowned safety science and human factors researcher at MIT. Her story is that two navy pilots ferry two aircraft to a different base, doing a training exercise along the way. Pilot B is going to fire a dummy missile. They load up the dummy missile and start their exercise. The pilot selects that missile tube and fires it. The airplane has a very simple autonomous system that notices when a firing tube has become occluded and selects a different one, so the pilot fires a live missile because the machine has no awareness that the context is not combat but training. That is where humans and autonomous machines working together can have many problems.

On the policy, I am very happy to see that the words "automation bias" have made it in there, in the context of attempting to reduce or eliminate automation bias. That is very difficult to do. There is an extremely active and long-running area of human factors research in how to reduce or eliminate automation bias. We do not know how. We are working on that and have some ways to start attacking the problem. Last week, I introduced a bug into my own code because ChatGPT gave me a bad review of my code and I accepted its change. Here I am giving evidence to this committee, and, although I know better, even I am subject to automation bias in my own profession. All we can do is acknowledge that automation bias exists and creates a risk when we build these autonomous systems. Eliminating it is probably a pipe dream.

Further on policy, I welcome the mention of the focus on time and space. Specifying the bounds of autonomy within time and space is extremely important, as is the boundary on target profiles. I would add that the sorts of context in which the weapons are used are important. Are we talking about an urban context or a battlefield away from civilians? Another thing about AI and autonomous weapons is testing. Are we talking about using weapons in a context and scenarios in which they have been well tested or something that differs from those training contexts? I will stop there because I think you are getting impatient, but there are huge problems related to that.

Taniel Yusef: I will try to be brief. We have had a good foundation from my colleagues and I think we will go into relevant detail with the latter questions. To begin, UK policy has shifted over the years. There is a misconception that there has been absolutely no movement. It is important to recognise that there has been some—not nearly enough, and not of the substantive kind that we would see as helpful. But there has been useful movement and it is important to recognise that.

To pick up on a couple of the relevant points that have been made, it is important to note that the CCW restricts us to state use of weapons in armed conflict and not municipal and domestic use. Given that developing weapons technology will inevitably drift into things such as domestic border control, or sometimes a state's use of torture on its own people, we all have to recognise that that happens and is inevitable, (without additional regulation). We will speak to that later. It is an important point to acknowledge. Some would say that it is another reason that CCW is not the appropriate place.

I direct us specifically to the phrase about "context-appropriate" human control; that language was used and quoted for us in detail. There are problems with that language. We prefer, and think more is captured in, the phrase "meaningful human control". You will hear it reiterated over and over again. There are technological and legal reasons why context-appropriate human control is not specific enough and does not capture enough detail. First, it fails to go into the detail of which contexts, what sort of board decides what is appropriate, how that is measured, who it is transparent to and how it is verified. There is so much missing within and

without that phrase that it is concerning. To be fair, I understand that the AI strategy tried to capture every bit of AI use from HR to autonomous weapons. One wonders what agility can be expected of from a proposal like that. It is asking an awful lot of a paper. I would recommend something completely separate for HR and for weapons. There is dual purpose and then there is asking for something to be that dynamic, which is too much.

As I said, if it is "context-appropriate", what is the context? Let us be more specific. I will try to keep this brief because, as I said, there is a lot more to go into and I do not want to take time away from much more relevant and detailed questions. There are situational contexts, as Laura went into. What kinds of battlefield? Are we talking about urban contexts, the sea or areas where there is an absence of human beings? I think we will go into the joint paper, where Article 5 unfortunately says something about warning people. That erodes the duty of states to take responsibility for their own actions, but rather puts it on civilians to essentially make themselves safe in warfare. That is not how laws of armed conflict work. You do not put the responsibility on civilians to get out of the way of war. There are problems with that element of context.

Then there are the legal and geopolitical contexts. Contexts can change in a split second. A negotiation happens at the UN and suddenly the context in which that weapon is being used is utterly different. You have heard others speak before me in more detail on this. The context in which an individual is a lawful target, whether or not they are a combatant, can change instantly. Context can mean something very different. The technological context can change, as can the coding. We will go into detail later and I do not want to over-speak. It serves us to go into some of the technological detail that I do not think has been presented enough to the committee. It would serve you to hear more about that.

On AI specifically, coding decisions and sensors in that more primitive autonomy, which you are aware of, have similar issues, but AI is more complex and problematic. There are human decisions. We must get away from the idea of there being a machine with an absence of human error because a human being is not in that process on the field. That is not the case at all. Human decisions are involved but not necessarily oversight. That is a different thing, which is why we emphasise meaningful human control. Human decision-making is littered all the way through the life cycle of the weapons system, from the capture and subjective classification of the data and what data is chosen to decision boundaries and parameters, labelling, coding and all the toggling that happens. We can go through that in a clear and simplified way, to really understand how just the arbitrary changing of a decision boundary can mean that something is captured within a target or is excluded; it could be that you miss an important target or exclude a protected target. You might exclude a child. That is not protected; you cannot see it. Either that means it is not protected and so is targeted in an explosion, or that you do not see a target that you need to present as an adversary.

Those are important decisions in coding and they are made by human beings all the way through the process. The myth presented, that there is no human error in it, would be to misrepresent the situation. Human error exists all the way through the process; so does computer error. Human beings fatigue on the battlefield; so do machines. You have to do the cost-benefit analysis. To think that one[i] is better than the other is a gross misrepresentation. When you talk about context, it is not just, "Is this a battle or not?" There are various contexts and one is technological.[ii] Who did the coding or the labelling? How many were involved? Not all labellers were created equal. The technology could be completely different if I did the labelling or Lord Sarfraz did it, and our thresholds could be meaningless. My 80% accuracy and his 80% accuracy might not be the same thing, which makes regulation quite different without a law on meaningful human control. I will leave it there.

The Chair: I realise how complex the issues we are dealing with are, but could you keep your answers as brief as possible as we have quite a lot of ground to cover in the time that is left.

Q158 **Lord Hamilton of Epsom:** I would like to focus on Stop Killer Robots, because I would like to know in slightly more detail what it stands for. I have a series of questions. Is it on a par with CND and would recommend that we took all autonomous weapons systems out of the British inventory, unilaterally disarming and setting an example to other countries that we hope might follow suit? That question is really for the two ladies.

Laura Nolan: I will start and Taniel can fill in any gaps. No, the policy is not that all autonomous weapons systems must be banned or disarmed. We are certainly not asking anybody to get rid of their missile defence and anti-rocket and mortar systems, for example. We stand for a lot of limitations and regulation. We favour a legally binding instrument to regulate such systems. In particular, the main question for regulation is whether such systems should be able to use humans as target profiles. We think that is very dangerous and problematic.

Taniel Yusef: Very simply, the answer is no, I do not think anyone is suggesting that we remove all autonomy. Autonomy and AI have their place. It would be an oversimplification to say that we should remove all of it. It can be a beneficial tool of technology, but there should be restrictions and limitations. On the way such regulation can be structured, there is an excellent paper by Article 36, which is part of the steering committee of the campaign. That sets out certain things that should be banned, such as the selecting, targeting and application of force against humans. That should be completely off the table for various reasons that we could go into. Humans are much harder to profile, and technologically it is different. It is unethical and difficult for legal reasons, such as ascertaining who would be a viable target as a matter of law.

It is technologically more difficult to target a human for all sorts of reasons. We are more agile, we move differently and we have different topography from, say, a big piece of metal. We are much harder to

target. Airport scanners find it difficult to target us, let alone something in the noise of war. There are all sorts of reasons why humans should be off the table, which we can go into. Legally, ethically and technologically, it is very difficult to target humans. There are technical reasons why it is hard but we go through gradations of things that could be controlled, with positive obligations and restrictions, so that not everything is off the table. Things come under certain controls. The short answer is no; I do not think anyone is pragmatically saying that.

Lord Hamilton of Epsom: In which case my next question is completely otiose. I thought Stop Killer Robots meant what it said on the tin. I am happy with that.

Q159 **Lord Grocott:** The Professor and Taniel Yusef said that there has been some movement in British policy, and seemed to say that in an approving way. If there is movement, it is towards something. What is the objective of the movement? What would give you supreme approval of British policy? What are you aiming at when you approve of what has been done so far?

Professor Christian Enemark: I suppose the movement I detect is that the UK Government, despite having a firm grip on being, in their own account, cautious about defining autonomous weapons systems, are apparently having that grip loosened by their allies in the UN system. That is reflected, for example, in the new draft articles document published last month. There, the UK Government have at least put their name to a prefatory statement about the subject of the draft articles, and that statement uses the same kind of language as the United States and the International Committee of the Red Cross have used in defining autonomous weapons systems. Albeit that it is covered by some heavily hedged language, I detect—perhaps because I am optimistic—that the UK Government might be, for diplomatic and pragmatic reasons, shifting to closer alignment with their allies on what autonomous systems are: on the core term that has to be put in place, broadly understood and subscribed to if any attempts to guide and regulate them are to have any effect. That is an example of the movement that I detect when I look at the documents.

Lord Grocott: It is a movement towards getting a definition. That is a start but it is not moving very far. Presumably, you want a definition with a view to that going further. I am new to this and am pretty astonished that there does not seem to be much agreement on precisely what we are talking about, which is a bit alarming. That is the approval, is it—that we are getting towards definition? What about you, Taniel Yusef?

Taniel Yusef: There can be movement away from things but in the UK sense, I do not think there has been.[iii] You can move left and right, and stay on the spot. There has been some helpful language, but I cannot recall a negotiation, whether on guiding principles or a treaty, which began with a definition. Negotiations of treaties have almost never begun with definitions. The language of definitions happens through the negotiation of a treaty. Having the impression that we must all agree, as

a starting position, would be a barrier to entry. It is rarely the case because, if we all agreed on it, it would not be so difficult. You move towards those terms.

It is difficult but that is one of the reasons why we need international law and a binding instrument to help agree the terms that people will be bound by in battle; that is, individual soldiers, because what happens is that you apply law to individuals when they use these technologies. You do not apply law to technology. The other thing about the UK attaching itself to a paper is that while the paper has good points, with some articulation that is similar to that of the ICRC, it fails to acknowledge things like targeting humans. Article 1 says, essentially, that illegal weapons are illegal. That is not helpful. As I said, another article puts the onus on civilians. That is not current law. It also fails the standard of UK language. They have assigned themselves to something that falls below UK standards, as articulated within the UN.

My personal judgment is that Russia and a couple of others have been very disruptive in international fora. It is across the board—not just in CCW but in every single one, and I partake in many others. That has included trying to disrupt and remove civil society participation. I am sure that other things are going on and I cannot speak for the minds of the UK delegation. Of course, I talk to them and they are very affable and bright people. There are other things going on and of course they want to partner with their allies. That is true, I imagine. Speaking in my personal capacity, as someone who has been in the diplomatic space and these fora for nearly a decade, my perception is that they are signing up to documents to keep something afloat while Russia tries to tear it apart.

Lord Hamilton of Epsom: You mentioned Russia, which we heard before is not co-operating on any of this. Is any international treaty viable if it does not include the Russians?

Taniel Yusef: Yes.

Lord Hamilton of Epsom: Why? Would a key component not be missing?

Taniel Yusef: To argue that no law works if another party is possibly likely to infringe on it would be to argue that all law is useless and we should return to a state of nature. I take the point; I am not trying to be dismissive. We are party to and uphold a number of treaties that any number of parties, including ourselves, could infringe upon. The point is that those laws exist not because we expect reciprocity but to set a standard. If those standards are not met, there are supposed to be repercussions. I think we would agree in this room that Russia has infringed upon international law as we speak and is committing a war crime. That has happened and there are repercussions. There may be repercussions into the future.

We do not set international standards because we expect everyone to meet them all the time. We set them as standards that must be met in

the future, by parties that may not exist yet, along with other standards. We must remember that this is not just about states; it is also about private actors who have technology that we do not have and will be leading the way. We do not create international law because we expect it always to be met. We create it as the standard that must be met. Also, we cannot accuse someone of breaking a law if we do not make the law first. If we do not have the law—

The Chair: I am keen to move on.

Q160 **The Lord Bishop of Coventry:** Thank you, Chair, and thank you, witnesses. This is a fascinating session, and my question takes us into an area where you have already brought us, in particular the draft articles from the GGE.

Professor Enemark, this sort of bridges the last question on British policy. You recommend that a concept of meaningful human control is built into defence AI policy. It sounds as if you are saying that these draft articles do not quite meet that criterion. I was very interested in the distinction that you made between agency-focused approaches and outcome-focused ones, as I think you put it. From my reading of what you said, I assume you think these are too outcome-focused rather than agency-focused. To broaden it out, on the basis that something like these draft articles can be improved and developed, I notice that Article 7 refers to their status. I am unfamiliar with how the articles then relate to law and are built into guidance or regulation, or what status that has. What are the mechanisms there? Initially, could we dig into the draft articles to see whether they get us anywhere close to the sort of principles of operation that our witnesses advocate?

Professor Christian Enemark: I will make a few remarks in response. I will not try to pick up everything—I probably would not be able to. You raised the very important point about an outcomes-oriented emphasis compared to an agency-oriented emphasis. Especially for techno-optimists who will say that AI is potentially a great way of overcoming the historically poor behaviour of humans when it comes to killing each other in war, the temptation is to say that, given that humans are historically so rotten across time and space at adhering to military ethics principles and international humanitarian law, surely it would not take too much at all to surpass that rotten standard with a kind of AI-generated system of creating effects.

The trouble with those kinds of outcome-oriented arguments is that all forms of consequentialist reasoning involve making predictions about the future, which is a shaky thing. Remember the wise words of Yogi Berra, manager of the New York Yankees: prediction is difficult, especially about the future. With consequentialist arguments, we are sometimes tempted to be too optimistic, even in the moment but also in what an individual category of enabling technology can give us. Instead, international humanitarian law and military ethics may be, and arguably should be, a whole system for ensuring that humans are in control of and take

responsibility for the use of force for political purposes. That is what war is.

The Lord Bishop of Coventry: Do you think the draft articles do not enshrine that sufficiently? Are you saying that they operate from that other place?

Professor Christian Enemark: The draft articles talk about things like commander and operator intention. You can detect elements in the draft article where there is clearly an emphasis on which agent we are dealing with. It is not an artificial moral agent but a human agent. There is a requirement that it is a human agent who does and decides things, and so forth. Bear in mind that these draft articles are merely intended to be an interpretation of existing law and do not have, as you suggest by mentioning Article 7, the status of law. Of course, they cannot be ignored by people interested in law because they could affect the way that we interpret existing law for better or for worse. I have my own gripes about another article in the document that maybe we can get to later.

Laura Nolan: These articles put some very difficult burdens on operators and commanders—for example, in anticipating the effects. Article 4 says that the commander has to anticipate the damage and whether the collateral damage will be excessive in relation to the attack. That is extremely difficult if the commander cannot choose or control the exact location, target, time and place of that attack. By the definition of autonomous weapons systems, that decision is happening in the weapon—in the machine. You are asking commanders to anticipate the effects of an attack that they do not fully control and cannot fully anticipate.

A core tenet of complex systems theory basically says that when you have systems with multiple components and actors interacting, and where the system has state, the number of potential outcomes grows exponentially very quickly, and it becomes very difficult to predict those effects. We see that already in the relatively limited degree of autonomy that existing weapons systems have—for example, with the well-known Patriot friendly-fire incidents from the Gulf War. We are asking people to predict something that is effectively unpredictable. In some ways, it is very concerning that an impossible task has been heaped on the shoulders of commanders. It is very difficult for them.

The Lord Bishop of Coventry: Thank you very much. Taniel Yusef, do you have anything to add? Is it the right approach to try to develop this sort of operational guidance? If it is, how does that make a difference?

Taniel Yusef: Without international law and an additional legally binding instrument, guidelines are merely self-referential and act on trust.

The Lord Bishop of Coventry: So you need something weightier.

Taniel Yusef: Yes. I have spoken a lot. I do not want to take additional time away from the future questions, which I think will dig into judgment

and proportionality, because there are technological reasons why that is so difficult.

The Chair: This moves very conveniently into the area that Lord Houghton of Richmond wants to explore.

Q161 **Lord Houghton of Richmond:** A common emerging theme of these evidence sessions is that the committee wants to make progress and the witnesses want to add complexity. I do not mean that as any form of criticism, but what the committee really wants to do is sensibly inform policy development in a way that does not deny operational advantage to national forces. In the context that, sadly, probably war is not going away and has not just started to be, but has always been, the realm of chaos, friction and uncertainty, this is just another element of it.

It is very apparent as we go through our evidence sessions that there is academic, scientific and specialist disagreement over the dangers of AI and the imminence of those dangers. There is some acceptance that the benefits of AI may actually eventually equal, if not exceed, the risks, certainly in the context, as the professor said, of a long history of human fallibility in the application of lethal force.

Could we therefore commonly agree that in policy terms—I think this is part of the underpinning policy of the UK Government on this—to proceed with caution represents an acceptable way forward, given the fact that this is a dynamic and not a static thing, because the technological advantage or current disadvantage is on a journey? That is as long as proceeding with caution is limited in two very significant ways: first, that you would never allow a field pass, or allow into service, a weapons system where the output of autonomy, or A, is outside human comprehension or cannot be forecast, and so cannot be given to a commander's ability or training to delegate to a machine; and, secondly, that all the permitting systems have their autonomy restricted by what we have been talking about, which is the retention of meaningful human control.

I get that that needs further definition, but those are effectively methods by which—I think you put this very nicely—within an autonomous system, the autonomy was not connected such that there were no fail-safes of human intervention so that the autonomous system could never in itself link a whole series of "inform, decide and execute" in a way that was not subject to the discriminatory and sentient superiority of a human being. You could allow, therefore, a gradual approach to whether such issues as distinction and proportionality can eventually be discharged autonomously, albeit that the legal responsibility would still be at the command level that authorised that autonomy. I hope I have linked that together.

Professor, perhaps you might have a stab at that. We want some common ground on which we can recommend progress rather than just adding to the complexity of what, to me, seems a somewhat academic stasis.

Professor Christian Enemark: I will have a go at the beginning and the end, if I may. I sense frustration, and you are very correct to say that this is fast moving. There is a war going on as we speak, and there is an appetite to crack on. I acknowledge that.

One of the first things you said, Lord Houghton, was that there was a concern not to deny operational advantage. That fits in with the concept of an AI arms race: if our enemy is racing ahead on this tech, we had better do that as well, otherwise we fear being outpaced and defeated, if it comes to that. The counterpoint is that you do not have to put operational advantage and caution in opposition to each other, either as a matter of military strategy in operation or as a matter of military ethics. Arguably—this particular space is a good example—caution on the technology is operationally advantageous, to the extent of having caution against mistakes such as would harm UK personnel or catastrophically harm civilians that UK personnel are charged with protecting.

Caution could be a way of avoiding hastily moving towards systems where hostilities are initiated or escalated at AI speed, to the detriment of UK interests, where things cannot be reined in because of technological factors. I mentioned the problem of fratricide—the simple safety of your own people because of particular technologies where an insufficiently cautious rollout has not been undertaken.

The final point that you made is—

Lord Houghton of Richmond: Forgive me, but you turned my comment into one where I was advocating an arms race, and you are sending back caution to me. I asked you to agree that "proceed with caution" is a fundamentally right approach: proceed with caution so as not to unnecessarily deny the military advantage that AI potentially has.

Professor Christian Enemark: I must have misunderstood.

Lord Houghton of Richmond: Dare I say it, the taking of a contrary or adversarial point of view about the concept of proceeding with caution appears to be the position of half the academics, professors and people who derive a living from this stuff. Anyway, keep going.

Professor Christian Enemark: At the risk of disagreeing with you again, I will pick up the last point. Forgive me, if you could.

The Chair: You are operating in a hazardous environment.

Professor Christian Enemark: In which case I will be very brief. We will need to return to this, all three of us, because you have put into the conversation in your final remarks earlier the concept of discrimination and proportionality being discharged autonomously. We really need to get into that idea. There is a philosophical and an ethical argument to be made that that is an impossibility. Only humans can do discrimination. Only humans can do proportionality. The autonomous discharging by a non-human entity is philosophical nonsense, arguably.

Lord Houghton of Richmond: Arguably, yes, because there are other communities in this who would say, "No, there is nothing the human can do that you can't put on a silicon chip".

Professor Christian Enemark: This is the argument. Let us have it.

Lord Houghton of Richmond: What I am saying is that you would proceed cautiously, and until that was absolutely proven you would not adapt that.

Professor Christian Enemark: It cannot be proven because I would argue that it is an impossibility.

Lord Houghton of Richmond: In which case we will never get to your problem set, but it still should not stop us proceeding cautiously.

The Lord Bishop of Coventry: Lord Chairman, may we clarify whether it is an impossibility on philosophical and ethical grounds that it could be reasonably done? Is it a priori simply impossible on philosophical grounds?

The Chair: Can we put Laura Nolan out of her misery as well? She is attracting our attention.

Laura Nolan: I am sorry to butt in. I have a small aside and a larger answer to the question. The small aside is that you mentioned human intervention to prevent things spiralling out of control. The human factors perspective on this is that humans are extremely poor supervisors of machine activity, particularly when that machine activity is fairly reliably correct. It is among the most boring things you can ask a human being to do. It is very difficult for people to maintain that engagement, so it is better to keep humans actively engaged in the process wherever possible. You will have better results.

My comment on your broader question about proceeding with caution is that we should understand that adding autonomy adds risk in all sorts of ways, and it is important to use it only when there is really an advantage. You do not want to sprinkle autonomy fairy dust on everything military. Where is the advantage in running autonomous attacks on a tank column? You would not want to do that. You would want to wait until it is in a ravine or somewhere that it would be advantageous to attack at that point. You want to choose your strategic timing in many cases. Use it in places where there is no alternative, where it is worth it and where the risks are manageable, particularly in relation to obligations under IHL. Then we could absolutely address your point.

The Lord Bishop of Coventry: I butted in, sorry.

Taniel Yusef: May I add a couple of things? Guiding principle (b) was agreed by states in the CCW: "Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system". That has already been agreed, in 2019.

Lord Houghton of Richmond: But we have systems already in service where that is not the case. Take the self-defence systems of certain ships. Because they have to operate at a speed that is faster than human decision-making, when certain forms of weapons systems inbound are detected, regulations on the battlefield will, in a certain context decided by humans, effectively delegate to that weapons system the authority to defend the ship.

Taniel Yusef: If that were, for example, the Phalanx[iv], it would be responding to a ballistic signature, which is not the same as an AI in battle updating and decision-making. They really are quite different things. Also, it is not a proportionality decision or a judgment where you are talking about collateral damage. If you were talking about proportionality and something like collateral damage, that can actually be debated by lawyers. You know what it is like to be in a room full of lawyers: they will have different accounts of whether or not something can be argued. There are different perspectives on that.

For technological reasons—we might come on to that in the women question—whether or not it has adequate data, it may have seen something or not. The data may be there, or it may not. Judgment has to be applied by a human being to say whether that is a viable decision. That requires human judgment and legal judgment, aside from the ethical question, so there is tech there.

To be a little bit defensive of my colleague, I think the point that he was making about caution was possibly a slightly different reading of your version of caution. I agree about proceeding with caution, but that is where we would say "have a legally binding instrument on use" in certain instances of use, where exercising the right kind of control and judgment within the law is simply not possible within the realms of predictability. There is a version of caution where you have to protect your own because of fratricide and all those other reasons—technological flaws, unpredictability and error. Every time you update something, you update the errors too. That is a fact. If something updates in the field it is no longer the weapon, after a certain point, that was reviewed in the Article 36 review. At what point are you still accountable for the same weapon? Those are just practical things. They are all issues.

I cannot speak for my colleagues, but I think this is an issue. We do not do enough things like red teaming. We do not do enough of that sort of preparedness for adversarial attack. This is a conversation that colleagues and I have. There is so much attention on the offensive, even though the offensive has all the issues of unpredictability and missing data—being unable to fill in the missing blanks of partial data. If a few pixels are missing, it is no longer a rifle; it is a 3D-turtle. We know that there are a plethora of examples. We do very little work beyond a box-ticking exercise on defending against the adversarial. That is a real issue. I think some of what my colleague may have meant about being cautious is part of having the operational advantage. Forgive me if I am over-speaking.

The Chair: Laura Nolan, do you want to add briefly to that?

Laura Nolan: Extremely briefly. I just want to caution against taking too many lessons from defensive-type systems, particularly missile defence and its friends. Those systems are typically collocated with human beings defensively. Human operators are there in the situation with the weapon. They have perfect context and can look out the window. That is extremely different from a system that is mobile and roaming around choosing its own targets. We need to bear in mind the huge increase in complexity.

Lord Houghton of Richmond: Do not get me wrong; I am not an idiot. I recognise the difference between installing those mechanisms on a ship and something that is highly complex in a built-up area among people. All I am saying is that, incrementally, it may be that advances in the technology will advance the envelope under which those sorts of delegations are safe.

Q162 **Lord Browne of Ladyton:** Chair, can I intervene and make a plea for some simplification of the question, even if not of the answer? I think I can confidently speak for everybody. We are not suggesting that we seek evidence in order to make recommendations that we should behave irresponsibly, or recommendations that will deny the necessity of red teaming and being very careful about all the stages of the process of artificial intelligence engaging in weapons systems. It is about making sure that we are not making fundamental mistakes or allowing errors to creep in that will have effects later which prevent us doing things that we want to do.

In that last question, we got very close to one of the issues that we are genuinely interested in, which is whether the current technical evidence suggests that algorithms trained by AI, or whatever, are capable of making the sorts of proportionality decisions that are necessary in a battlefield. That is the question we are really interested in. There is a philosophical and moral issue to this, but if you enter that relatively simple question through the philosophical gateway you never actually get to having the question answered.

What we would really like to know from you is your understanding of the way that algorithms can be trained and developed, so that they are capable of the level of proportionality that we demand, in particular of the sorts of lethal decisions by the people who are responsible for them in the battlefield, which we cannot delegate to machines. I do not think any of us is looking to find an argument to allow us to do that if it is not technically possible. If we could get some response to that, it might be very helpful, otherwise we get into a lot of areas in which we have already taken yes for our answer. We do not need to be persuaded that some of these things are utterly dangerous. None of us, I think, is actively looking for evidence to support putting them into our report.

Laura Nolan: I can start very briefly. First off, I do not think that proportionality is that simple. What you are trying to do with proportionality—

Lord Browne of Ladyton: No, I am not. I am telling you that we know the rest of the context. We are seeking advice about this particular technical position. We are short of evidence about the technical capability, so that is what we are after. We have acres of evidence about the rest of it.

Laura Nolan: If you are asking for a proportionality judgment, you need to know the anticipated strategic military value of the action, and there is no way that a weapon can know that. A weapon is in the field looking at perhaps some images and doing some machine-learning perception stuff. It does not know anything. It is just doing some calculations that do not really bear any relation to the military value. Only the commander can know the military value because the military value of a particular attack is not purely based on that local context on the ground; it is the broader strategic context. It is absolutely impossible to ask a weapon on the ground to make that determination.

Lord Browne of Ladyton: Thank you.

Taniel Yusef: Can I give a brief example of how an algorithm would work in the field? I gave this presentation in the UN recently. An IHL lawyer for the MoD of another country approached me and said, "I didn't know that is how maths was applied in these weapons". I presented a support vector method. This is very fundamental. The words do not have to mean anything. I could present another algorithm; this one is quite basic. It classifies data points and finds what we call a hyperplane, which is just a line. It draws a line between data points, which is basically a scatter graph—just dots on a graph.

To be really clear, if it were pixels or an image, it is not seeing that. It is not seeing a me—a Taniel. It is scattering dots around as pixels on a graph. According to a mathematical line that was drawn by an individual—a coder who has created that line, which is a mathematical equation based on probabilities; it can be a straight line, a curve or round, and its various lines overlapped can look like a fried egg—the classes are essentially where the pixels will be allocated between those lines according to statistics. I am oversimplifying a little, but that is essentially what happens. Where pixels are allocated between the lines will be where they are allocated according to that class, which becomes the identification.

There is a nice graph that I kindly stole from GitHub, because I thought it did this much better than I could. I put all up these graphs and I said, "This is what this other version might do". It had weights and parameters and pixel dimension. It had a cat at the beginning—because we like our cat/dog example, don't we?—and it did a calculation. It put the pixel allocation through its mathematical formula, and at the end, essentially, the number of pixels that were in the highest outcome went into the "right" class. If it scored 150 as opposed to 100 or 90, it was allocated to that class. The image had cat, dog, ship. It could have been Reaper drone, aircraft carrier, woodlouse. It could be anything, because the machine does not know what these things are; it is allocating classes and

deciphering those pixels. It is a bit more complicated than that but it is just maths, and quite rudimentary. But because of the computational methods and the power, it is doing maths very quickly. That is all. It is very simple maths as well; in the UK, you learn it when you are about nine, 10 or 11. So the highest number goes to that class, and it decided this cat was a dog. It was not a dog; it was a cat.[v]

What concerns me is that when that happens in the field you will have people on the ground saying, "These civilians were killed", and you will have a report by the weapon that feeds back, "But look at the maths". It is not just false positives. It will be a recording from the field that says, "But the maths says it was a target that was a military base". I might be simplifying—but not misspeaking—but that is essentially what it is. It will say, "It was a military base because the coded maths says so". We defer to maths a lot because maths is very specific, and the maths will not be wrong; it will be right. There is a difference between "correct" and "accurate". There is a difference between "precise" and "accurate". The maths will be right because it was coded right, but it will not be right on the ground. That terrifies me because without a legally binding instrument enshrining that kind of meaningful human control, with oversight at the use-end, that is what will be missing.

You asked the question, Lord Browne, about proportionality and if it is technically possible. No, it is not technically possible because you cannot know the outcome of an AI modelled weapon system and how it will achieve the goal that you code it until it has done it, and you do not know how it has got there.

Lord Sarfraz: The other day I saw a dog that I thought was a cat. It happens with humans as well.

Taniel Yusef: I assume you did not shoot it.

The Chair: Lord Mitchell, you have a question on procurement and industry.

Q163 **Lord Mitchell:** This is probably an easier question to answer. I think we can push along quickly on it. What do you see as the role of private industry in contributing to and influencing the debate on AWS? As a subsidiary question, is there a widespread unwillingness among AI developers to work in defence and national security? What implications does that have for the implementation of the Government's defence AI strategy?

Laura Nolan: What should be the role of industry? Certainly, industry will have opinions. You should be very conscious that a lot of industry they are trying to sell you something, as I am sure you are well aware. Again, I am not saying anyone is stupid here but take their opinions with a pinch of salt. Many manufacturers of weapons have voiced an opinion that clearer regulation on these matters would be welcome, because they do not know what they can build responsibly, ethically or legally. One piece from industry would be more clarity. Having more clarity on the

bounds of how far nations will push towards AI weapons and autonomous weapons would be helpful in reassuring people working in that industry that their work will not be turned towards ends with which they are not comfortable. I will leave it there.

Professor Christian Enemark: The role of private industry is, first, to help Governments make the world safe for business, and, secondly, to do what private companies do, which is to maintain a commercially valuable reputation for trustworthiness among their customers, employees and shareholders. I do not have any secret insights on the exact level of willingness or unwillingness by AI engineers to go and work for Governments or for defence industry firms, but I have heard debates among such professionals about whether one ought to do that.

You can look at the lessons of history to say that there is money to be made by companies by taking on some big defence contracts, sure, but that can sometimes be very risky, especially when it comes to commercial reputation, if your company is not already in the defence industry. Most tech companies are not already in that space. The level of willingness to contribute to government efforts in this space could quickly drop off, for example, if there was a scandal or an incident in which such and such a company's technology was found to have been involved or implicated. Companies will not rush headlong into playing along. They will be very concerned, in thinking long term about reputation, about what their involvement might be.

Lord Mitchell: Is there much mutiny by staff?

Laura Nolan: I left Google because it was insisting on pursuing Project Maven, so, yes, there are certainly some of us.

Taniel Yusef: I reiterate that the role of private industry for the most part is to pursue profit, which is understandable as companies have to pay their staff. There is a lovely quote from Alain Tremblay, who said: "The restriction on the weapon system is a client decision, not what we're providing. If a client says 'I'm comfortable from a rule-of-engagement, a law-of-armed-conflict and Geneva-Convention (angle), to have this thing finding a target and engaging a target'—it can do that". Similarly, German industry a couple of years ago said, "We need international law to provide the lines and the standards. It's not our job to do it". The CEO of Saab at the REAIM conference, I think just a few months ago, said, "We are putting in our own lines because international law and states have failed to do it".

My concern is about leaving this in the hands of private industry—that is not because companies cannot be trusted. Some indeed cannot, while some are wonderful and provide excellent goods, services and jobs. But how many instances have we seen where companies have been charged with incredible malfeasance yet have covered it up and not been forthcoming with corrections? If they are in charge of this sort of thing, what will be the relationship thereafter? What would be the relationship with government in trying to pursue some kind of remedy?

Proceed with caution but bearing in mind, as everyone here knows, that the technology is very much led by private companies. This is dual use and dual purpose. You are not going to be able to force companies to hand over IP in this regard, because they are, rightly, going to protect it. We live in a porous multinational company world, where you have trade and knowledge being exchanged everywhere, including a couple of salespeople who I have heard in this room. I would be conscious of those who have not wanted to turn up here, which I understand.

When companies oversell the capacity of the technology, which I think is misrepresentative, and say things like— and I have discussed this with tech colleagues—"It is immoral not to use it", as I have seen in other areas of this technology such as driverless cars, which are now shown to have incredibly high death rates comparably, when you look at the real data, they do not present the real flaws of the technology or talk about how they would regulate its use, safeguard or test it. They say misleading things like, "We'll test it". To a technologist, "We'll test it", is the beginning of a question; it is not the answer to one.

Q164 **Lord Sarfraz:** I have a technical question about datasets. This is all about having very high-quality datasets, right? Is enough work being done on auditing datasets?

Laura Nolan: To the best of my knowledge, there are not many military-specific datasets of any size and such datasets that exist are private to the militaries that have collected them. People talk about datasets all the time. It is not clear to me that standardised datasets, such as faces of people and English text, are particularly helpful in a military capacity. To build weapons, you would need combat datasets, which, as I said, do not really exist. Given the happily relative paucity of combat situations, they will be hard to collect.

Lord Sarfraz: What about non-combat datasets—just datasets? Is there enough auditing happening of datasets? Are there enough third parties, big auditors and so on, who are auditing datasets?

Laura Nolan: It is a great question. To my knowledge, auditing these datasets for things such as bias, which I assume is what we are talking about, tends to be done by individual researchers. There is an Irish lady who has done some work like that on image datasets at UCD, but it is patchwork.

Lord Sarfraz: Why are the big auditing firms not entering the space of auditing datasets?

Laura Nolan: Nobody is paying them to do it. And they probably do not have the skills.

The Chair: Lord Triesman, you are interested in this area. Do you want to come in with a supplementary?

Lord Triesman: Lord Mitchell has asked all the questions that I wanted to ask. Thank you, Lord Chairman.

The Chair: Thank you very much. I have been increasingly remembering the line from Andrew Marvell about "Time's wingèd chariot hurrying near". I think it has actually caught up with us. We have really enjoyed having you with us. You have opened our eyes to a whole lot of perspectives, shall I say, that had not quite arisen in our inquiry to date, so we are all really grateful to you. Thank you very much indeed.

---

[i] Note by witness: This specifically refers to the assertion that computers would make better analysis than humans- meanwhile the computer error can be exponential when parameters set by humans and deeply imperfect data necessarily predate battle-implementation.

[ii] Note by witness: It is important to note that different algorithms will present different analysis of the same data and the same algorithm with different parameters will have different outcomes. This will lead to varying battle images and decisions. Sometimes contradictory ones, requiring additional human judgement. There is no ground truth in AI. Statistics can be selectively chosen for a desired outcome, presenting additional problems.

[iii] Note by witness: The UK delegation have done some very good work within CCW emphasising the UK's support of legal accountability throughout the chain of command, repeatedly. They have also elevated language around restricting AI use within time and space.

[iv] Note by witness: It is important to note that the Phalanx already has significant human oversight, is restricted within time and space, responds to a specific ballistic signature and does not update its own target profile but even so has mistakenly, fatally shot down passenger airlines. As AI integration evolves, this human oversight is an increasingly important factor there. For instance, see the downing of an Iranian passenger jet during the Iran-Iraq War.

[v] Note by witness: In such cases, at the most basic identification stage, the classification is wrong. This can easily happen, for example, when data is "over fit" and is tested very well for the background noise. This is essential testing for battle but can mean that the AI identifies the noise better than the object. Again, we cannot see this until it manifests poorly enough in the wild. It will test well in the lab but only degrade in the field, misleading at procurement stage.