HOUSE OF LORDS

Artificial Intelligence in Weapons Systems Committee

Corrected oral evidence: Artificial intelligence in weapons systems

Thursday 27 April 2023

10 am

Watch the meeting

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

Evidence Session No. 4 Heard in Public Questions 43 - 63

Witnesses

I: Professor Mariarosaria Taddeo, Associate Professor, Oxford Internet Institute; Dr Alexander Blanchard, Digital Ethics Research Fellow, Alan Turing Institute; Verity Coyle, Senior Campaigner/Adviser, Amnesty International.

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

Professor Mariarosaria Taddeo, Dr Alexander Blanchard and Verity Coyle.

Q43 **The Chair:** Good morning and welcome. Thank you very much indeed for coming and sharing your expertise with us. I hope that we will be able to wrap this session up by about 11.30 am, which, I am sure, will give us plenty of time to explore the issues that the committee is particularly interested in. To begin with, could you introduce yourselves and your origins in this area?

Verity Coyle: I am a senior campaigner and advisor at Amnesty International. I work for the international secretariat on military, security and policing issues. We are a member of the Campaign to Stop Killer Robots.

Professor Mariarosaria Taddeo: Good morning. I am associate professor and senior research fellow at the Oxford Internet Institute at the University of Oxford, and DSTL ethics fellow at the Alan Turing Institute. I am a philosopher and have been working for the past 20 years on the ethics of digital technologies and how this reflection should inform governance.

Dr Alexander Blanchard: I am the DSTL digital ethics fellow at the Alan Turing Institute. By background, I am a political theorist and I work on the ethical implications of using AI in defence and security.

The Chair: What is the distinct character of artificial intelligence in weapons systems that gives rise to a need for additional concern or scrutiny? I am using AI there not just in the narrow sense of AWS—completely independent AI-powered systems—but as introduced in any stage of the deployment of a weapons system.

Professor Mariarosaria Taddeo: We need to take a step back here, because it is important to understand that, when we talk about artificial intelligence, we are not just talking about a new tool like any other digital technology. It is a form of agency. It is some kind of artefact that can interact autonomously within the environment and learn from this interaction.

This is the root, so to speak, of all the advantages we have when we use AI, but also the origin of all the ethical issues and risks we face when we use this technology in any domain. When we apply this technology to the use of force, whether directly or indirectly, there are three sets of issues that make it very problematic.

The first one is the limited predictability of the outcomes of AI systems, which is much lower than with other technologies. The predictability is a result of technical features. Some models of AI are less predictable than others, but it is a constant characteristic.

It is also a consequence of the context in which we implement these technologies. Some contexts and environments might have a prompt that

we did not foresee at the time of deployment, which might trigger a new, unpredicted and unwanted behaviour by the systems.

The last element is adversarial. AI technologies are very vulnerable and fragile, and can be relatively easily manipulated by third-party actors, which might prompt the systems to behave differently from what we expect. This is the first issue.

The second issue has to do with responsibility. Attributing responsibility is a key element when we wage war. It is a fundamental aspect to make sure that we wage war while maintaining its morality. This was stated in the Nuremberg trials, for example. Attributing responsibility means that we can trace an outcome back to the intention of the actors who performed those actions.

Because of this limited predictability, we cannot really do that when it comes to AI. It is very problematic to say, "This action was intended by the designer, the developer or the user", and this creates a responsibility gap that is very hard to fill. On top of this is the way in which we design and develop AI, which is through a very diverse and distributed net of expert engineers that might change over time. If we are going to use AI in warfare, we have to make sure that we can apply regulations, including with respect to the responsibilities that people have, which is very hard to do and we are very far from finding a solution.

The last issue is that AI systems perpetrate mistakes, as humans do, but they do so very effectively and in a much higher range than humans. We can imagine bias, as we have learned about it in the past 10 years, being applied to the use of force, and so targets being identified for the wrong reasons. These are three macro issues that are very relevant.

The Chair: Do I take from what you are saying that it is our lack of knowledge at this stage as to how a system might operate that is leading to this high level of caution and that, to turn that coin over, if we had better knowledge of how systems operated, we would not need to be so cautious in the terms you have expressed it?

Professor Mariarosaria Taddeo: I would say that it is quite the contrary. It is about having very good knowledge of the limits of this technology, which will not be overcome or solved by better knowledge of the systems. The predictability issues are intrinsic to the technology itself. It is unfeasible to imagine that we could overcome them, so it is more about our knowledge of the structural issues of the technology.

Verity Coyle: We do not believe that the only measure of danger in autonomous weapons systems is their lethality. An autonomous swarm that is entirely non-lethal but immobilises people and automates surveillance can be highly oppressive and violate a range of human rights. We see the negotiation of a legally binding instrument as a correct response to potentially catastrophic threat posed by autonomous weapons systems with AI within them.

The use of AWS, whether in armed conflict or in peacetime, implicates and threatens to undermine fundamental elements of international human rights law, including the right to life, the right to remedy and the principle of human dignity. The threat to human dignity posed by delegating life and death determinations also presents ethical problems and raises concerns under the principle of humanity under IHL's Martens clause.

AWS raise other serious human rights concerns outside of situations of armed conflict, threatening the right to life, the prohibition of torture and other cruel, inhumane or degrading treatment or punishment, and the right to security of person. We are currently undergoing a technological revolution in many fields driven by AI, machine learning, miniaturisation and automation. This includes computer vision, sensor-based environmental analysis enabling facial recognition technologies, and complex autonomous action in the world.

The use of these technologies poses risks in many fields, but the risks are particularly acute in autonomous weapons systems that can use force against targets, potentially automating killing. The distinction between these weapons systems and many other types of conventional weapons boils down to the question of meaningful human control over the use of force. Without this essential ingredient, AWS cannot be used in compliance with IHL and IHRL.

As we have seen across many domains, artificial intelligence and machine learning can, as the previous speaker said, introduce unpredictability, opacity and bias in systems, where systems are making life and death decisions, and this certainly violates fundamental principles of IHL and IHRL.

Over the last decade, many states, the ICRC, campaigning NGOs, academics and roboticists have been debating these issues, in particular how these systems can be regulated in order to maintain meaningful human control over the critical functions of selecting and attacking targets. While these debates touch on many familiar IHL issues, such as indiscriminate effects and the principles of distinction and proportionality, they are qualitatively different from the previous debates about the regulation of non-AI weapons and deal with profound issues around human agency and control, digital dehumanisation and fundamental accountability gaps.

The Chair: We will probably come back to the question of the practicality of binding regulation, which you touched on at the beginning of your intervention. Dr Blanchard, do you have something to add to that?

Dr Alexander Blanchard: I do not have much to add beyond what the two previous speakers have just said. The comments that I prepared parallel Mariarosaria's quite closely. There is an additional issue, in that, as AI systems become more pervasive in their use by the military, one potential problem is understanding how different systems interact with one another and the behaviours that that might generate.

Q44 **Lord Grocott:** Verity, going back to this wretched business of definition, which we have hit on a number of occasions, for the Campaign to Stop Killer Robots, what is a killer robot? More specifically, can anyone on the panel describe a system in operation somewhere in the world at the moment that is an example of a killer robot? The question is about specific examples rather than hypotheticals, and I would really like to hear from the other two witnesses as well.

Verity Coyle: We would define a killer robot as a system that uses sensors to determine when and where force will occur. We would not categorise it as a particular type of weaponry, but more a process whereby information is gathered via sensors, a calculation is made as to whether this matches a previously determined target profile, and then force is applied without any human intervention in that step.

We are talking about a process. In terms of regulation, that is the right way to go, because, as you asked for a current example of a killer robot in operation, things are developing at pace. We are seeing more and more systems being operated with an increasing set of autonomy within them. It will not be long before all these parts of the process come together and can be deployed.

IHL is not sufficient because these types of combinations of systems were not in place when IHL was defined, and we now need to create new law to govern the new systems and processes that we are looking at.

Lord Grocott: That is very helpful. You are saying that these things are in the process of being developed, although they have not got there yet. In that case, can you give me a specific example of something that is nearly there?

Verity Coyle: The closest example would be the Kargu-2 drones deployed by Turkey. They have autonomous functions that can be turned on and off. It is a case of whether a switch is used to stop the autonomous function being fully autonomous or whether there is still a human in the loop. We are on a razor's edge in terms of how close we are to these systems being operational and deadly.

Q45 **Lord Hamilton of Epsom:** My question is a follow-on to Verity. We have Phalanx working on Royal Navy ships, which work without human intervention and intervene when they are attacked by rockets or aircraft. That is an example of AWS working today. Would you scrap that, so that the Royal Navy could not use that weapons system any more?

Verity Coyle: This brings us on to the point that we were heading towards in terms of internationally legally binding law. We have a couple of recommendations in that field, which I think will answer your question. We would like a robust, well-implemented, legally binding instrument, which must include a positive obligation to maintain meaningful human control over the use of force and a ban on certain systems.

Those are systems that cannot be used with meaningful human control and, most importantly when we are talking about the UK Government's

position, have a human as a target. There should be a complete ban on autonomous weapons systems being able to target human beings. In that instance, we are talking about a different use of weaponry.

Lord Hamilton of Epsom: If a Royal Navy ship was being attacked by an aircraft, it would have a human being in it.

Verity Coyle: Missile control systems are in place already and operate under IHL. If it is in that parameter, that is a system that is already within use.

The Chair: So your answer is that you would ban it.

Verity Coyle: If it is targeting humans, yes.

Professor Mariarosaria Taddeo: There is an interesting question about the definition, which is one of the recommendations that I wanted to bring up at this committee. At the moment, we lack an internationally agreed definition of autonomous weapons systems. This is very problematic, for two reasons.

First, the definitions that have been provided so far by different state actors have been done so cynically. Very high thresholds for the requirements of autonomous weapons systems—almost sci-fi requirements—have been set. This has done nothing but create confusion, a sense in the public that this topic is not being discussed seriously at a state level, and space in which actors are free to design, develop and test these weapons without having to call them autonomous weapons systems. This is very problematic.

The second element where the lack of a definition is problematic is exactly this kind of discussion. We focus on the autonomy of these systems, but it is not just that. We have plenty of automatic weapons in place already, although some are not allowed, such as landmines. It is not just the autonomy that is a problem there. It is the learning ability and the adaptive behaviour of these systems that make it so problematic.

One thing to do is to start thinking about a definition that is realistic, that is technologically and scientifically grounded, and on which we can find agreement in international fora to start thinking about how to regulate these weapons. This is very important.

The previous speaker highlighted a very important topic. Autonomous weapons systems are not sci-fi. We are not looking at something that is going to come in five, 20 or 50 years. We have seen them already deployed. What makes the difference is whether those weapons have been used in the capacity or capability of being fully autonomous or whether there is a human in the loop. The war in Ukraine has defied the taboo of autonomous weapons systems, and that is why it is important that we move on to identifying these weapons and start thinking about regulating them concretely.

Dr Alexander Blanchard: In terms of whether the systems have been used, the United Nations identified, in a report, Kargu-2 by STM as the first use of an autonomous weapons system. As Verity also highlights, the problem is that it is not clear whether these are being used in an autonomous setting. What tends to happen is that an arms manufacturer creates a system and markets it as fully autonomous, because this is good for sales. It gets used and draws publicity, and the manufacturer then changes the specifications that it announced around that system to say, "This isn't actually a fully autonomous system. It still has levels of human control".

There is a real question there not only of definition but of transparency of the use. I am not sure how you address that, but it is about knowing whether this system was indeed used as a fully autonomous, or autonomous, weapons system.

Q46 **Baroness Hodgson of Abinger:** My question is on security. What must be done to safeguard the use of AI in weapons systems both from potential outside interference by agents with bad intent and from inherent problems and malfunctions?

Verity Coyle: Systems that cannot adequately predict, control or explain their effects should be banned. Examples are systems that use machine learning to change their mission critical parameters independently, or ones whose target profiles, period of time activation and geographical scope of operation are too broad to allow for predictable outcomes.

Other systems such as certain missile defence systems must be subject to rigorous and frequent testing and oversight to ensure that they can be deployed predictably and reliably, with sufficiently narrowly defined profiles of military objects, geographical scope and activation periods.

It is not really our role to talk about outside interference in this space, so I will pass to my fellow panellists.

Professor Mariarosaria Taddeo: There is first a question about the mindset. We should realise that, with AI, there is no way in which we can rule out those risks for good, because it is a learning technology, it is unpredictable and it is inherently fragile or vulnerable.

There are ways in which we could improve the robustness of systems by making sure that the systems will continue to behave as we expect it to, even under adversarial or different circumstances than what we considered at the moment of design. We have to understand that these measures come with a cost that is not only economic but organisational. The deployment of AI sometimes requires a restructuring of the way in which we address issues like security, safety and control. This is very important, because it allows us to understand that the security of AI comes in the remit of the governance of AI, and we can do something in that area.

For example, one way of improving the robustness of AI systems is to enforce adversarial training, where you have different systems sparring

against each other. This is a well-known technique, but, because AI improves itself by feedback loops, the way in which it interacts with another agent allows it to change its variables and coefficients.

We could imagine certification processes and standards that mandate this form of training to make systems more robust than they were at the moment of design. We could also imagine governance that is more granular and mandates standards for this kind of training, so that we can set higher thresholds before defining a system as robust.

The other element goes back to control and monitoring. Ex ante control with AI is no longer sufficient. One way in which AI is attacked is by adversarial actors interfering with the system and making it change its behaviour just enough for it to behave favourably to the attacker, but not enough that the legitimate users would realise it in time. This is a crucial issue, because it means that we lose control without realising it.

One way of avoiding this problem is to imagine deploying a system in the wild and a cloned system under controlled circumstances, and using the second system as a baseline. Whenever there is a gap or too much of a variance between the system in the wild and the one deployed under controlled circumstances, a flag is raised. Of course, this is costly in terms of time and resources, but, in high-risk contexts such as defence, is crucial. The risk is just too high not to be considered in this way. I hope that this addresses your question.

Dr Alexander Blanchard: A number of defence institutions have identified traceability as an important part of designing, developing and deploying autonomous weapons systems. They identified traceability as part of understanding the system. If we can trace the parts of the system, we will understand it when we come to use it.

However, traceability may have an additional value, in that, if you can trace where the components of the AI system have come from—who has a remit over them, who has supplied them and who has designed them—you also have a process by which you may be able to secure the supply chain behind the model, so to speak. This will become important because of the fact that lots of actors may be involved in designing and building an AI system. From creating and collecting the data and building the algorithms, you need to have a sense of what organisations have been involved in the design and development of those systems.

Baroness Hodgson of Abinger: Would you ever see it being necessary to embed some kind of system where it could be neutralised by the people who have made it, to stop it being taken over?

Professor Mariarosaria Taddeo: Overridability and a kill switch is crucial. We could not imagine relinquishing that sort of control. By default, there has to be a way in which a human can intervene and stop the machine from working. That is essential. I do not think that they have been developed in such a way that this is not possible.

Q47 **Baroness Doocey:** Professor Taddeo, I want to make sure that I understand what you have just said. You said that it is easy for third-party actors to manipulate automated weapons. You then suggested that, if I was sending something into the field to destroy another thing—not a person—the best way to make sure that nobody could interfere with it without me realising would be to send out a second AI system to make sure that it hit the target or, if it did not, that I would understand that the first one had been manipulated. Is that what you were saying?

Professor Mariarosaria Taddeo: Perhaps I was not too clear. It is more that you might imagine, when sending any kind of robotic equipment equipped with AI into the wild, there are hundreds, if not thousands, of parameters that you want to control or that need to function. One of those parameters might be manipulated by a third actor, making the system behave in a different way.

It is easier to know whether one of those parameters has been tampered with if we have a testing condition in which we can observe the same system doing the same operation, not in the wild but in an environment that we control, so as to know what those parameters should be as a baseline, and then compare those two things in order to see whether the second one is being manipulated.

For some weapons, knowing that the system has been manipulated once it has been deployed does not make any difference, because you cannot intervene once it has been deployed, but it might make a difference for the next iteration if we know that the second weapon that we might want to deploy has been tampered with.

Baroness Doocey: You are saying that this should just be done in a testing environment.

Professor Mariarosaria Taddeo: There is one testing environment and one wild environment. What I am saying is that we should keep monitoring the way in which the same system behaves in the two conditions, so that, if the system deployed in the wild behaves very differently—and that is the threshold to be set—we will know that something is off and we can intervene before too many mistakes are made.

Baroness Doocey: I understand what you are saying. I am not sure that I agree with it.

Q48 **Lord Browne of Ladyton:** I am much more interested in stopping people from interfering with our systems than I am in finding out who they are or who was responsible for planting the problem in the first place. I am an avid reader of reports from the Defense Science Board of the Department of Defense, which have convinced me over about 10 years that it is impossible to protect any system that is connected to the internet from interference. Should we not disconnect all of this stuff from the internet so that it cannot be interfered with, given the level of risk?

Dr Alexander Blanchard: Which aspect do you want to disconnect from the internet?

Lord Browne of Ladyton: I am delighted that we have you to give evidence, but we need technical people who can tell us how to do this, not why it should be done or retrospectively. It is easy to say that every component, even if it is made in a sweatshop somewhere, has to be properly audited to make sure that there has not been some algorithm put into it that makes the machine work in a particular way. It is almost impossible to do. If we are going to allow machines that think for themselves to be affected by people who can get access to them, should we allow that to be a possibility at all?

Dr Alexander Blanchard: This seems like a question for all weapons systems in general.

Lord Browne of Ladyton: It is a question for all weapons systems, but this is an escalation of risks.

Q49 **Lord Houghton of Richmond:** You have to put a boundary on the ethics of this. Quite a lot of extant military capability in the field of electronic warfare is all about messing about with the technology of other nations' or an enemy's capability. Electronic warfare is used to jam things and to stop things flying. Chaff is used to decoy missiles and send them to other places where the enemy does know they are going to be. If you are not careful, you run the risk here of making all of warfare unethical, because it is naturally both a human and a technological struggle.

Do not get me wrong. I am also on the side of those who want to do away with killer robots, but I would just go about it by maximising our exploitation of AI. I just wonder why on earth you have not, therefore, visited electronic warfare as a fundamentally disturbing element of war from an ethical point of view, because you are, by and large, ethical in your approach to this.

The Chair: Professoressa, that was probably aimed at you, but anyone else can come in.

Professor Mariarosaria Taddeo: Having spent 15 years writing about cyberwarfare, I have focused a lot on electronic warfare and cyberwarfare, and the ethical implications that those new technologies bring about.

Going back to the specific question, there is an advantage in having AI used online, because it allows the systems to learn, to interact with the environment and to adapt their responses. There is no binary answer to your question. It is a cost-benefit analysis. It is even more convenient in some contexts to have it online. Is there a risk that they will be manipulated or attacked? This is an ethical question, because the cost-benefit analysis implies some analysis of the ethical risks, of the justness of war and of the human rights that we want to respect.

There is no one-off answer to this question, because you have to know about the context, but you point to a very important issue, which is exactly the need to consider the different elements coming in.

Going back to electronic warfare, tampering and sabotage is an element not only of warfare but of international relations in some cases. The whole point is that, when it comes to AI, we have limited control over the effect that that might lead to. We might sometimes find ourselves in a situation where we have a very unwanted outcome and no culprit, because the sabotage was not intended to cause those outcomes, the scale and effects are too big and too wide, and we are not able to ascribe responsibility for those outcomes in a just and fair way.

This is the difference when we factor AI in. Whether it is in cyber or kinetic, it is the same issue, but AI brings a new set of aspects that are qualitatively different from the ones before.

Q50 **Lord Sarfraz:** Verity, is there a path that you can see where, when we look back five years from now, we can say, "AI and autonomous weapons systems coming along has made the world a safer, better and brighter place"?

Verity Coyle: No.

The Chair: That is the shortest answer that we have had this morning.

Verity Coyle: If I may, though, there is a path where the world can act responsibly in this manner. Last October, over 70 states at the UN General Assembly put together a joint statement. They came from different positions on this issue, because it is complex. There is a defence, security and technological advancement question. The UK wants to be part of all those conversations and play a leading role. However, the UN Secretary-General, thousands of technologists of very high standing, and all of the human rights and humanitarian organisations believe that you need to act in a manner that creates new law to govern this threat.

This allows me to come back to the Phalanx question. I misspoke. That system operates through a very defined time and space parameter. It is able to be switched on and off. It does not learn and it does not change its parameters as it goes. That would come under the positive obligations that we are seeking in this legal instrument, not a complete ban. But my simple answer is no.

Q51 **Lord Mitchell:** This is the first time that I have addressed this Committee, so I must declare that I have no conflict of interest. The question is to do with intelligence analysis. What are the challenges and concerns associated with the use of AI for intelligence analysis that may later inform the use of a weapons system?

Dr Alexander Blanchard: I would just put a quick question back to Lord Mitchell. What do you mean by "inform the use"?

Lord Mitchell: I am not sure that I can answer that.

The Chair: We are interested in the way in which a corpus of intelligence can be deployed in order to make a weapons system effective or more effective, or to decide when to deploy it. That is a related question.

Dr Alexander Blanchard: I undertook some work with Professor Taddeo on the ethical issues raised by the use of AI for intelligence analysis. We did a review of the literature, within which we identified several factors that potentially introduced problems for using AI for intelligence analysis. They were to do with disproportionate levels of intrusion into personal data, problems with explainability around the decisions that come about from using AI, issues of bias in the decisions that come from the AI system, issues of authoritarianism, and problems with collaboration and classification.

I can see that a lot of those apply to national security institutions and the intelligence community. They would have application to defence organisations, in so far as they undertake intelligence analysis as part of planning.

How close is the link between the system and the intelligence analysis you are using? I would assume—and I would certainly hope—that, when a military organisation decides to deploy an autonomous weapons systems, the intelligence that it is using as part of that deployment has been rigorously checked and accounted for, that there has been red-teaming of that intelligence, which is to say that it has been challenged, and that the usual procedures and processes for checking the veracity of a piece of intelligence are in motion.

I cannot envisage a situation, although I may be totally wrong, where intelligence that is generated using AI is then being used to inform the deployment of an autonomous weapons system without first checking that intelligence. There may be issues where a piece of intelligence is used and it turns out that that piece of intelligence was wrong, and then you have to go back and interrogate the AI system to find out why it was wrong. This is why I asked what you meant by "inform the use of a weapons system".

The Chair: What you have described appears to be something that might be happening in rather a relaxed timeframe, whereas we are talking about something that would develop very quickly indeed.

Dr Alexander Blanchard: In that case, the problems that I identified earlier on—questions of bias, wrong decisions and not having meaningful human interpretation of the intelligence being produced—would apply, given a situation where intelligence analysis is feeding straight into the weapons system.

The Chair: Professoressa, do you agree with Dr Blanchard?

Professor Mariarosaria Taddeo: Yes, very much. We worked on this topic for quite a long time. There is one addition to what Alex just said, which is that there is an element of tech bias whenever we embed AI into

decision-making processes. The human, if not properly trained and if left without internal policy assurances that they can disagree with what the machine is suggesting, might be inclined to say what the machine says. It is the approach of, "The computer says no".

Under stressful conditions, this can be very problematic, because the outcome of the AI system is acquired uncritically and quite dogmatically. One element that we have to make sure to consider is how to embed AI into decision-making processes, whether or not leading to kinetic outcomes. Doing so means training the personnel appropriately and having the appropriate safeguards to protect human autonomy—the ability to disagree and to do otherwise.

This is, again, a line of work that requires investment and cost. We have to be able to shoulder those economic and organisational elements in order to limit that risk.

Verity Coyle: Inasmuch as the intelligence analysis informs targeting decisions, all the same issues would apply. Is there meaningful human control over the system? Are the outputs predictable and reliable? What biases are baked into the system—for example, a reliance on under-representative data training sets? Is the system auditable? If it is impossible to understand how certain results have been produced, what happens when things go wrong? Who would be held accountable for a catastrophic miscalculation based on faulty outputs from an inherently opaque and perhaps biased system?

The Chair: Earlier on, I said that we would certainly come back to the question of practicability of regulation, and Lord Fairfax of Cameron is about to prove me right.

Q52 **Lord Fairfax of Cameron:** I had a question about the possible effectiveness or otherwise of international regulation in this area. At the moment, if you talk about Ukraine, there are allegations of tens of thousands of war crimes, but whether those will ever be successfully prosecuted I do not know. This is of concern in the AI area.

I was then going to go on to the question of proof. We were talking about AI weapons systems and proving that a totally autonomous weapons system that has no human involvement has produced a bad outcome. Is that irrelevant? All we are talking about is the outcome. If you have a good outcome, no one is going to complain. It is only when you have a bad outcome that you are going to look at the situation and try to determine whether the system used was entirely autonomous. I would welcome any comments about the problems of proof, but you may think that it is irrelevant. Sorry, there are a few questions there, but the first one is about the possible effectiveness of international regulation.

Verity Coyle: You could ask that about any international process or, indeed, international rules-based order that the UK professes to support. No international norm is universally complied with, but all set benchmarks and can stigmatise non-compliance. We have seen this with the conventions on chemical weapons, anti-personnel landmines and

cluster munitions. Ratification and adherence to these conventions has never been universal, but all have consolidated norms and driven positive change.

The goal is a long-term one of driving out the most problematic systems, so that their use is marginalised to a handful of wrong actors, and of promoting and strengthening overall compliance with IHL and IHRL. Proliferation is an issue with any weapons system, but the creation of a new legally binding instrument can set up systems for monitoring, information sharing and best practice, which can curb unlawful development and transfers.

The UK Government should be leading these initiatives and setting out principled arguments to strengthen international law in this area. Given the pace of technological change, the creation of new international law is urgent.

Professor Mariarosaria Taddeo: As the previous speaker said, regulation is the one issue where autonomous weapons systems do not differ from any other weapons. As Lord Fairfax just mentioned, Ukraine is a good example. We are, unfortunately, seeing routine breaches of international humanitarian laws, but that does not lead us to question the validity of those laws. It is only because those laws are there that we can imagine, one day, holding the villains to account.

The point with laws, if you do not mind me wearing my philosopher's hat again for a moment, is that the international community states, recognises and commits to respect the fundamental values and rights of our societies. These are statements and commitments that go both outwards, so internationally, and inwards, where the state commits to its citizens, "When waging war, we will not behave in a dishonourable way, but do so in a justified and just way". It is a way of building trust internally.

If AI is the mark of the way in which digital societies are going to wage war going forward, and if we agree that the way in which societies wage war is a mark of the level of civilisation and the values that they respect, we cannot let those who are our opponents set the standards, because we cannot follow them in a de-escalation of regulation. Those laws are there to protect our values, whether or not AI is involved, so it is crucial that we work to define them.

What is problematic with autonomous weapons systems is that we have had a very polarised debate for decades, which has not allowed any progress in any possible regulation, and so we now find ourselves with these kinds of weapons being used with an astonishing regulatory gap and no ability to hold anyone to account for mistakes or errors.

You asked a question about proof, if I understand correctly.

Lord Fairfax of Cameron: It is about evidence, really.

Professor Mariarosaria Taddeo: It is hard to understand what element of an AI system might have caused a mistake or an unwanted outcome, but I hope that it is not hard to trace where that mistake came from, in the sense of whether it was a human or a weapon. I would imagine that there is a strong traceability process for whatever decision is made in war, so that we know that that weapon caused that mistake.

We might not be able to explain why that weapon caused that outcome, and that is problematic, but we want to be able to at least trace it back to the technology. Otherwise we have an issue that is not to do with the use of AI but with the way in which we keep records of war actions, which will be very problematic in terms of accountability and responsibility.

Q53 **The Chair:** We have dealt with the regulation of nuclear weapons for an extended period, but the creation of weapons-grade plutonium involves very heavy industrial processes that can easily be identified and monitored. We are talking now about weapons systems that might be non-compliant and in the hands of people we would not like to see possessing them, but they can be put together in an anonymous shed in some suburb. That is giving you a real problem in terms of regulation, is it not?

Professor Mariarosaria Taddeo: I come from a place where there is a very high level of burglaries, and it is very hard to catch burglars who come to houses and commit this crime. We did not stop locking our doors for that reason. It is the same thing. It is not really true that anybody can assemble an autonomous weapon that can cause proper damage and compete with something that a state might have.

It is also true that we are less likely to trace those weapons if we compare them to normal weaponry or to nuclear weapons, but this cannot be an excuse to stop aiming to regulate and trying to do so rigorously. What happens when we catch people who are assembling weapons that they are not supposed to if there is no regulation that says that behaviour is illegal?

Q54 **Lord Sarfraz:** As you said, Lord Chair, there is a difference between the development and the use of capabilities. If we overregulate the technological development for ourselves, and our adversaries do not play ball, we can get left behind very quickly. Is there a risk of slowing down our potential to develop in the process?

Professor Mariarosaria Taddeo: It is also a question about how we enforce this regulation and whether we can incentivise adherence on a scale that makes regulation more effective. Regulations are not just to be drafted or imposed. When it comes to international relations, they are also a way of creating and incentivising consensus on regulation with other means. That is another way of addressing your point. I am quite convinced that we should not let those who do not respect fundamental human rights or democratic values set the standards for the regulation of this technology, which is very dangerous and very risky.

Dr Alexander Blanchard: Just to add to the point about using it, you need brakes on a car to be able to drive it well. A technology that is regulated is probably also one that is governable and usable. In the context of the debate about autonomous weapons systems, people often say, "What about China?" Well, what about China? We have our own values and our sense of what we want to do, background democratic norms, how we wish to use tech and how we wish to perform on the international stage. If someone says "jump", you do not just jump. You think about how you wish to adhere to your own values.

I do not want to talk about regulation development here, but regulation around use. It does not bring any military advantage that I can conceive of to use a system in a way that does not comply with international law. There is no military advantage in using a system that cannot respect the principle of distinction between combatants and non-combatants. I do not see what military advantage a system that does not do that would give you.

Q55 **Lord Hamilton of Epsom:** Verity Coyle has said that we need new international law. I think she would accept that there is not much point in new international law unless the major players are involved—China, Russia, the United States and NATO members for a start. What evidence do you have that there is any consensus among these major players to get involved in new international law? Incidentally, this was called for by the previous report on AI by the House of Lords in 2018, since which time—and correct me if I am wrong—absolutely nothing has changed.

Verity Coyle: I am sorry to disagree with you, Lord Hamilton, but things have changed. I have been attending the international discussions at the United Nations for several years now. I have sat with and discussed this topic with colleagues from China, America and most NATO countries. There is a real willingness to work towards regulation. On some parts, they are looking at soft law—non-binding guidance—which is the route that the US takes in most of these processes, but does end up with a form of regulation that is helpful.

It still falls short of what our campaign and my organisation would call for, but it is moving in the right direction and has developed in both thinking and policy coherence quite dramatically. The UK's position has evolved slightly. It still does not go as far as we think it should. What we are seeing now with the UN statement from over 70 states last October is a growing head of steam towards new international regulation.

I was in Costa Rica at a meeting of all the Latin American and Caribbean Governments earlier this year, where the Belén communiqué was developed and signed by those Governments. It calls for a legally binding instrument. There will be meetings in several regions of the world this year outside of the CCW process, which will be setting the pace and the standard for this work. The UK risks being left behind, although it is very much wanted in those discussions. It was invited to the Costa Rican multiregional event but did not attend. We will also see the opportunity

this year at the United Nations for that group of 70 to potentially go further and to look for more action outside of the CCW.

One problem with the CCW is that it does not include all states. It is a smaller group of countries, but these weapons will affect everybody, and so everybody should have a voice in how they are regulated. It is correct and right that this discussion moves to the United Nations in New York, where all member states can discuss and make progress.

Q56 **The Chair:** The tension that we are all grappling with is that it is a given that the development of AWS is dizzyingly fast, but that progress along the road to effective international regulation is glacially slow, and it seems very difficult to resolve that tension.

Verity Coyle: I had the privilege of working on the international Arms Trade Treaty as an advocate and lobbyist for nearly 10 years. It feels slow, but we are taking all the steps that we need to in the AWS regulation debate. The UK Government were not onboard with the Arms Trade Treaty to begin with. Eventually, there was a moment where the change happened through committees such as yours making recommendations, and they were a leader and a very strong advocate for how that treaty was put together.

The United States, Russia and China all took part in those negotiations. They may not have signed up and be adhering to the principles, but we are stigmatising things around the world through the creation of new international law, and we must do it with this issue too. It is just a matter of time before the UK is on the right side of this.

The Chair: The lambent phrase in your answer was "nearly 10 years".

Q57 **Lord Fairfax of Cameron:** I want to ask one question that has just occurred to me listening to you all. It is about the doctrine of self-defence under the UN charter. Does it or could it trump concerns about totally autonomous weapons systems? In other words, would it be justifiable for a state, defending itself against an attack by another state, to deploy totally autonomous weapons systems in self-defence?

Dr Alexander Blanchard: You are asking, if a state is being attacked, whether it is okay under the doctrine of self-defence to use a system that is not traceable.

Lord Fairfax of Cameron: No, it is not traceability I am asking about, but a totally autonomous weapons system. State A is under attack from state B. State A deploys totally autonomous weapons systems to defend itself against the attack from state B. What do you think about that? Let us assume it is desirable that totally autonomous weapons systems, as everyone seems to be agreeing, are outlawed. Would the doctrine of self-defence trump that?

Dr Alexander Blanchard: I am not a lawyer, so I will not give you the legal answer.

Lord Fairfax of Cameron: I am just interested in your opinion.

Dr Alexander Blanchard: A lot of the work we do looking at the ethics of using autonomous weapons systems is against the backdrop or framework of the just war theory, which is this very adaptable and flexible, but also well-grounded and historically founded, ethics system. The point is that having a good reason to go to war does not give you carte blanche on how you conduct that war. If you have a just cause for war, that does not allow you to override the principles that determine just conduct in war. It does not allow you to ride roughshod over principles of distinction or proportionality. You are still bound by the principles that determine just conduct in war.

Q58 **Baroness Hodgson of Abinger:** I very briefly wanted to pick up on what you were saying. Why do you think the UK is not more at the forefront in pushing for these regulations? As you say, in the end they were very robust about the Arms Trade Treaty and chemical weapons. Why do you think the UK is lagging?

Verity Coyle: I speak as Amnesty, so we can talk about what we have seen and what has been demonstrated in the meetings. There is a block, predominantly by Russia, on consensus at the moment. There is no opportunity within the CCW for the states, even if they wanted to, to go much further. Even the chair's reports are being stripped of any content and no progress is being made on paper, even though much rich discussion and policy coherence is taking place.

The UK finds itself in a point of tension, given its stated commitment to be a technological leader in the world and to have innovation, tech and business move forward. We do not see the two things as being incompatible, but I feel that is one of the major blocks for the UK at the moment.

If regulation is in place, it will help that because the lines will be clear about what can be developed and harnessed, and what should not be put out into the world, either on the battlefield or in the hands of our police forces. The other problem with the CCW is that it is restricting this debate to military use but, if we look at any weapon development, we see that everything that starts in the military eventually ends up in police forces and on the streets in front of civilians. Tear gas and kinetic impact projectiles were never designed to be used in the ways that they are now, and we see their use, in many cases, amounting to torture and other ill treatment. We would like to have that regulatory framework in place.

The UK is acting constructively in the discussions at the CCW, and it has also been very positive in ensuring civil society is able to take part in those discissions, because some Governments are attempting to block access and to close those discussions down totally. It is appreciated that the UK has vocally taken the stance of continuing to include civil society even when we do not agree.

Q59 **Lord Houghton of Richmond:** My question really tries to anchor us back in the practicalities of how we exploit artificial intelligence in weapons systems, because, after all, what the committee wants to be able to do, at the end of all these evidence sessions, is to come up with some recommendations, which will help inform our own national policy and, beyond that, international policy. The question is what specific mechanisms, technical or procedural, are needed to enable accountability and transparency in the use of AI in weapons systems.

I am a professional soldier by background. Beyond the generic definitions when we are talking about AI in weapons systems, it seems to me there are probably three areas where AI can be sensibly used. One is to inform a decision; the second is to actually make the decision; the third is to carry out that decision.

To give a rough example of how that would apply, as was touched on in the intelligence question, you would draw on some element of AI capabilities—it might be digital voice recognition and facial recognition within a time-sensitive targeting environment—to give the intelligence certainty or degree of certainty that the thing you were going to attack was the legitimate target. In many respects, the AI is used to inform the human decision and there is sensible human judgmental involvement of a targeting board, ultimately chaired, potentially, by a Minister or a senior officer. That is all taken care of.

A human then empowers AI to act, which might be something relatively simple. For example, you launch a complex weapon with a fancy nose that is preloaded with target data, which can electronically recognise and fix into a target because of its electronic signature or whatever. Again, there is a human being who has determined to press the trigger, even though the ultimate weapon effect is delivered because of a clever piece of AI.

The third is the one we are most concerned about, when the artificial intelligence informs the artificial intelligence. The whole thing is done in a way where there is not, as we keep coming back to, this meaningful human control. As we have already touched on in various meetings, and in this one, there are already systems in place where effectively the human decision has been made to authorise the system to operate. The artificial intelligence, therefore, is the intelligence that makes the decision and fires the weapon, for example in the self-defence of a ship, because, dare I say it, the target is in technical terms pretty simple to discriminate and the time factor involved in making the decision is so critical that it makes sense. The legal responsibility still goes to the person who has delegated authority to an autonomous system, so we have not surrendered that, but to all intents and purposes you have allowed artificial intelligence to create the intelligence, make the decision and carry out the effect.

If this is a useful set of ways of defining artificial intelligence, there is nothing particular in that that goes beyond the current scope of regulation within the laws and frameworks of the battlefield as I know it. I have said to friends on this committee that the battlefield is a hugely

regulated environment. Through targeting directives, target delegations, rules of engagement, it is set up to be, with the addition of good training, compliant with IHL and all that.

Is there a need for a different framework because of AI, or do we simply adapt and exploit the current framework? Patently, some of the things you are worried about, the killer robots, would not make it on to the battlefield. They would not be permitted. They would be regulated out at a higher level. Going back to the question on specific mechanisms and technical procedures to enable accountability and transparency for the use of AI in weapons systems, can you suggest things that are helpful?

The Chair: I think Lord Houghton has set you a substantial essay question. Let us see how you distil your answers.

Professor Mariarosaria Taddeo: Let us see whether we can grapple with it. It is a very interesting question. Indeed, the issues we are discussing today refer to a context where an artificial agent is left to identify, select and attack a target. The AI then informs the AI, which informs the AI. Different levels of risk in terms of ethics emerge with context. You might imagine in submarine warfare an AI system, a fully autonomous weapons system, where you do not have such high-level risks of indiscriminate effects or breaching the just war theory principles or international humanitarian law, because of the context. The context you describe is one such example. The conditions are very much controlled and, while the risks are there, they have a very low probability, because the factors that might interfere with the machine are very limited. We have a control exerted on the environment.

There are other contexts where this is not the case. The war we see, for example, in Ukraine is waged in cities and streets. The environment is much more complex and there are variables there that might impact the way these machine behave. This might lead to outcomes we do not want to transpire.

Going back to your question, in that case, when something goes wrong, as was said before, what remains problematic with AI is to do two things. To ascribe responsibility, not just legal responsibility, but some kind of responsibility that we have to be able to ascribe to individuals qua individuals, is very hard for the reasons I gave at the beginning of this session to do with AI systems. If we do not have so much control, we run a very high risk of indiscriminate effects of these weapons when they are used in a complex environment.

These are the issues that make the use of AI in warfare very problematic. I should say, I am not dogmatically against these weapons. I am open to listen to conditions and to find compromises, but we have to be very much aware that the risks are concrete. A scenario we might find ourselves in is an officer being asked to take responsibility for something that they do not control, that they do not understand and that might put them in a horrifying situation. This is the situation we are looking at.

Going back to the question about how we improve accountability and transparency, that is the thing with this technology. We do not have full accountability or full transparency. It is always a question of trade-off. Sometimes, for example, transparency comes at the cost of the efficacy and efficiency of these machines. It is a balance. The question there calls for a technical answer. We have to set thresholds and quantitative measures for this requirement; we have to translate this high-level concept of transparency and traceability into specific requirements that enable it to be implemented.

A machine is more or less transparent depending on the model that we use and on the control that we have on the database that we use to train it. All these factors need to be regulated, because they drive risks and possible harms. Mine is not a specific answer, but it is time we sit down with engineers and we specify requirements and thresholds for these requirements, because this way we can understand how much risk as societies we are happy to take for this machine not to be transparent, since there is a chance it will behave a little differently than we expect. That is the not-so-ethical answer, perhaps, at the end.

Lord Houghton of Richmond: I am concerned with the fielded battlefield weapons systems. You do not have to cross the threshold of conventional warfare; you have weapons systems in low-intensity conflicts, peacekeeping operations and such. You can take an awful lot of that problem away by having some external regulatory boards. They exist, although I do not know the proper names of them—something 38 or 36. When those weapons systems do not sit within a threshold that permits proper human involvement, but what they are called upon to do in technical terms has been proven by repeat testing to have a fail-safe record of 99.9%—there will always be some sort of fraction—you remove the problem from the battlefield commanders about the judgmental use of certain weapons systems that are wholly AI-enabled. That, to me, is a far more sensible way forward, or else the training load and judgmental load that you are piling on to those actually fighting the wars are too great.

Professor Mariarosaria Taddeo: I agree. The point is that we still lack the criteria to make these decisions for autonomous weapons systems. We do not have it at this stage. There is still a level of decisions that officers will make on the ground, given the scope of weapons they can use. If those weapons include AI, they will have to have an understanding of the technology they are deploying. Some gap of understanding and knowledge still needs to be closed, even if we regulate and define the set of capabilities that can be put in the hands of those people. Ex ante, with the board, there are serious difficulties in deciding the criteria, deciding the threshold and making sure, as we might touch on later, the review is effective for these technologies.

Dr Alexander Blanchard: As part of your initial question, there was a question there about whether we are using existing frameworks or completely new ones. Putting aside the need for new regulation, and the

extent and scope of that, there will be a lot of adapting and evolving existing frameworks to take account of the novel capabilities that AI introduces. To give you an example, we just mentioned Article 36.

Lord Houghton of Richmond: I did not really know what it was, but I know it is to review weapons systems.

Dr Alexander Blanchard: We could flow into that. The issue with Article 36 is this. Say you are undertaking a review of a weapons system. If it is an AI system, it may undergo upgrades in a way that introduces certain functions and capabilities that take it beyond the initial review. It is not a question about whether Article 36 still works, but about how you adapt reviews under Article 36 to take account of the way in which AI systems can transform.

The United States Department of Defense has raised this question explicitly in its refresh of its directive on autonomy in weapons systems. What do we do when an AI system refreshes itself, evolves or transforms in a certain way and upgrades itself? The DoD has said that, when there is a substantial change in the weapons system, we have to undertake a new internal review of that system, but it does not go on to qualify what "substantial" is in that context. That is part of the issue here. How do you get more granular? What is "substantial"?

There may be witnesses who can talk about this better than I can, but article 36 tells you not that something is lawful, only that it is not inherently unlawful. Whether its use is lawful it is still dependent on context and how you use it. A contextually appropriate use of this system is still important here and it is not going to be one size fits all. How do you determine certain measures around the use of that system to keep it lawful or ethical?

In the example you gave about an AI system that is responding to an incoming projectile on a ship, chances are, if it is metal and moving at 600 miles per hour, the AI system is going to identify the right thing. Similarly, if you have an autonomous sub at the bottom of the ocean and it comes cross something big, metal and beeping, there are probably certain assurances you can have about what that is. There is not going to be a one-size-fits-all thing here and it will be about adapting existing frameworks.

The Chair: I am afraid the clock is rather against us at the moment, so I would like to go to Lord Hamilton to follow up specifically on that Article 36 issue, and then come to Lord Browne and the Lord Bishop, who have some questions about ethical dimensions.

Q60 **Lord Hamilton of Epsom:** Lord Chair, I think perhaps the business of Article 36 might have been answered by that question. I would like to come back, if I could, to Verity Coyle, because I am still slightly smarting from her accusation that I was completely wrong on the whole question of nothing having happened. She said, if you remember, that there were frightfully productive conversations going on with the Chinese and,

indeed, with the United States, but she subsequently said that the Russians were being extremely difficult about all of this. It seemed that the Russians had a veto over any progress in terms of new international law. Is that correct? Do we have to wait for the Russians to lift their veto before there will be any new international law?

Verity Coyle: No, and I am sorry if you feel smarting, Lord Hamilton. It is a real privilege to be here talking about these issues with you all.

In the convention on certain conventional weapons, everybody has a veto through the means of consensus. If one state raises an objection, a report cannot be agreed or language cannot be put forward into the official record. When I say it has been blocked, it has been blocked in that particular forum where fully autonomous weapons systems are being discussed. There are many other forums developing concrete political and official working level opportunities for progress to be made.

In the CCW, the Brazilian chairperson is doing a fantastic job of bringing the states together, trying to make progress and thinking about what the questions are that countries need to answer together. Unfortunately, we end up with empty reports because at that point a state will use its veto and its consensus opportunity to block. I would not want us to think that is the only issue, though. There are other states that have not had to put their lack of desire to move forward completely on the table because they are shadowed, at the moment, by the position taken by that state.

Lord Hamilton of Epsom: We do need unanimity.

Verity Coyle: It depends where we are going to have this discussion, so, no, not in essence. That is part of the reason why the CCW is failing to make progress. At the previous review conference—these happen with high contracting parties every five years—there was a real hope that a negotiating mandate would be achieved within that group. Unfortunately, that was not the case and, as we get closer to the next review conference than to the previous one, countries, particularly those in leadership positions on this, are thinking, "Let's move forward to somewhere where we know we can make progress", hence the change of forum is really important now.

The Chair: We might have to reflect on the difference between consensus and unanimity. We can do that.

Q61 **Lord Browne of Ladyton:** Thank you to all three of you for your evidence. This question, at least the initial question in this group of questions, I want to direct quite deliberately to you, Professor Taddeo, because we want to take advantage of the fact you are a member of the Ministry of Defence AI ethics advisory panel and give you an opportunity to share your experience. Just before I ask you the specific question, I warn you that I have, as will be obvious, an informed bias about the area of algorithms.

This advisory panel was referred to in the policy paper *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-enabled Capability in*

Defence, which is from 15 June 2022, so now approaching a year out of date. There is an annexe referring to this panel that you are a member of, having been invited and no doubt accepted. You did not design the panel, so you are not responsible for that.

It tells us in the executive summary that the panel advises the second Permanent Secretary "on the development of policy relating to safe and responsible development and use of AI". He, indeed, is a member of the panel. It then goes on to say, "As of the date of publication, the current panel has met three times" and then surprisingly to me it says, "The panel has not been involved in the creation of policy related to lethal autonomous weapons systems, nor the department's policy on AI safety". These seem to me to be contradictory sentences. I do not know whether you agree with that, but if they are advising, and the Permanent Secretary, who is responsible for this, is part of seeking this advice, it seems to me by definition you do have some impact on the creation of the policy.

Never mind that. It is the membership of the panel that really interests me. This is a hybrid panel. It has a number of people whose interests are very obvious; it has academics, where the interests are not nearly as clearly obvious, if they have them; and it has some people in industry, who may well have interests.

What are the qualifications to be a member and what is the process you went through to become a member? At any time were you asked about interests? For example, are there academics on this panel who have been funded by the Ministry of Defence or government to do research? That would be of interest to people. Where is the transparency? This panel has met three times by June 2022. I have no idea how often it has met, because I cannot find anything about what was said at it or who said it. I am less interested in who said it, but it would appear there is no transparency at all about what ethical advice was actually shared.

As an ethicist, are you comfortable about being in a panel of this nature, which is such an important element of the judgment we will have to take as to the tolerance of our society, in light of our values, for the deployment of these weapons systems? Should it be done in this hybrid, complex way, without any transparency as to who is giving the advice, what the advice is and what effect it has had on what comes out in this policy document?

The Chair: I should make it clear that all witnesses in front of us are entirely protected by parliamentary privilege.

Professor Mariarosaria Taddeo: I hope not to use it. It is a good question. I was not involved in the process of selecting the members of the panel. I have a clear, but not secret, conflict of interest, in so far as part of my research is funded by DSTL, which is a research branch of the MoD. These are very good questions, which the MoD should address.

I can answer part of your questions in saying that the panel was brought together because this was part of the defence AI strategy and included some principles on ethics of AI. It was understood that there is a need for

expertise in considering what the ethical questions are and how this should be addressed. It was put together in good faith in trying to understand these questions and not leaving them to practitioners rather than academic experts.

In my mind, it is a good thing that this is a multistakeholder panel, because any panel that has to discuss issues where legitimate but conflicting interests are involved needs to have representation of all those interests, as long as there is a balance, and autonomy over the decisions or the discussion is guaranteed, which is the case in this context. There can be improvement in terms of transparency of the processes, notes and records. I agree with you, and this is mentioned whenever we meet.

Aside from the specific processes that led to the establishment of this panel, there is one element of advantage. A panel such as this is very much needed in any defence organisation, because there is a tendency otherwise to flatten ethics on security or safety measures, or to devolve ethical responsibilities to practitioners, which might not have the required understanding to address the multiple trade-offs and balances that applying and thinking about ethical questions imposes.

This discussion is one hour and a half, and there are a lot of experts in the room who are all prepared, but we did not even scratch the surface of many issues that we have to address. Such panels are very much needed. I agree with you that the process and the procedure could be improved, but it is also important that we keep in mind that this is a crucial element.

I would also like to stress that this is an advisory panel. We might give feedback on texts such as the principles you mentioned for the ambitious, safe and responsible use of AI. It is a step in the right direction, but it is the first one. We should welcome the idea of having ethical panels, which also have a veto or review power over the decisions that are made, and can check the adherence of the organisation to the principles it has given itself.

Aside from the procedural issues, which I agree are there, I think this is a good effort in the right direction. I would hope it is not deemed sufficient to ensure ethical behaviour of defence organisations; more should be done. I do not think it should be in the remit of this panel; there should be other panels and boards where ethics has stronger leverage, so to speak.

Q62 **The Lord Bishop of Coventry:** Following on from that, as you say, it is such a good thing that the ethics committee is there. Like Lord Browne, I am fascinated by how it actually works.

Turning back to Lord Houghton’s question on how AI intelligence is applied to battlefield situations, what power do members of the group have to inform, empower and implement? How seriously do you find yourselves taken, and not just in discussing interesting and fundamentally important things? Some fascinating written evidence has

come in, which has said very positive things about the ethics committee, and particularly the ethical principles that our Ministry of Defence is working with, but has also said that, while they are very good as principles of intention, the next step, as you are implying, is how they translate into policy in practice. I am interested in whether you see any evidence of that happening: the application of ethics, under the influence of this ethical committee, to policymaking and practice.

Professor Mariarosaria Taddeo: I would distinguish between the two remits. The panel is an advisory panel. So far, all we have done is to be provided with a draft of, for example, the principles or the document and to give feedback. It was very specific feedback in my case, for example on wording, using some words or mentioning some things rather than others. The feedback has so far been taken very seriously. I also note the limits of the scope, in that we did not talk about autonomous weapons systems.

I do not see that as a contradiction, in so far as the panel was put together to do a specific job at the time, which was to give advice on the drafting of those principles. It was a different scope or remit. It does not mean it will not happen in the future, perhaps, but it was a way of creating a safe space where the debate on the ethics of AI in defence, which is much broader than autonomous weapons systems, could occur without being monopolised or drifting completely towards this topic. There are many ethical issues that are as severe as the ones we discuss today, but have nothing to do with autonomous weapons systems. This is one level of answer on the functioning of the board.

The panel is not responsible for the application of the principles, and it should not be, because it is an advisory panel. That is what I was referring to. There should be, in my mind, some other board put together to oversee or lead efforts on translating the principles into practice. From what I see, there is concrete interest and effort being made to operate this translation, and it is a point of particular delicacy. If left only to practitioners, the temptation could be to translate those principles—which resemble constitutional principles in their nature, being very high level and in plain language—into simple operational measures, missing the need to balance those principles against each other in some circumstances, or missing the point that those principles require ethical and critical reflections on how to better implement and interpret the spirit in specific contexts.

This is a particular point to which it is important to draw attention. There is a need for expertise to drive this translation, and it is not necessarily an expertise that sits within any Ministry of Defence. It requires leveraging expertise that is in society, and that is why I welcome the idea of a board, whether it is the ethics advisory panel or another panel. Importantly, it would be a panel with a different remit, a panel with a stick, so to speak, that can veto or review operations, which is not in the remit of this panel.

Lord Browne of Ladyton: Is there value in having an independent,

transparent body in the ethical area to review the implementation of these principles and to report on them, so that we can all know whether the application of them is consistent with the tolerance of our society to live within our values? It seems to me a simple enough question.

Professor Mariarosaria Taddeo: The answer is yes.

Lord Browne of Ladyton: I am interested in what the other two panellists have to say about this. You implied that.

Verity Coyle: We would always value transparency of committees such as this, but for the reasons you have outlined I do not have information about what they have been discussing or the workplan. I will leave it there.

Dr Alexander Blanchard: Yes.

Professor Mariarosaria Taddeo: If I may, that is not what this board does.

Lord Browne of Ladyton: I understand that now.

Professor Mariarosaria Taddeo: None the less, it is important that the ethics advisory panel is held to higher standards of transparency and accountability. Just for clarity, this is not what the board does. It does not have any veto; it is not driving any effort to translate the principles into practice. For the sake of completeness, we need not only a board that oversees the way the Ministry of Defence behaves in terms of ethical behaviour. We need expertise in a board to show how to translate those principles into practice and into criteria. That board should also be put in place and should be held to high standards of transparency and accountability.

The Chair: We are pretty clear on the division into two functions there, which has been brought out by Lord Browne's question. I am in danger of leaving Lord Sarfraz out unless he can distil his question into something very brief indeed.

Q63 **Lord Sarfraz:** The AI regulation White Paper has five principles, and it says that in different domains regulators should come up with their own thinking. Do you agree with that? Is there alignment between that White Paper and the defence AI strategy in the *Ambitious, Safe, Responsible* policy document?

Dr Alexander Blanchard: I would give a short answer to that. Both *Ambitious, Safe, Responsible* and the White Paper leave a lot of space for those who have a particular understanding of a given domain or application of AI to work out how those principles are applied. However, that is the beginning of the work. All the work has to follow on from that. How do you bring those principles down to that more granular application within the expertise that a given regulator will have. *Ambitious, Safe, Responsible* leaves a lot to be done, as does the White Paper, although my expertise on the White Paper is not as deep as others'.

Verity Coyle: We agree with some of the risks outlined—in other words, algorithmic bias, responsibility and accountability, unpredictability, unintended consequences and incentives, and potential lack of human control. The UK Government's opposition to the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifestyle is a step in the right direction, and the broad idea is good. However, we would not agree with the specific language, especially "involvement" rather than "control". It should also emphasise lack of meaningful human control as a key defining feature of prohibited weapons systems.

Professor Mariarosaria Taddeo: In terms of the approach proposed in the White Paper, being domain dependent and domain-centred is very good, because users of AI in different domains pose different risks, so a one-size-fits-all approach will not work. However, there are a few elements that are very problematic in my mind. The White Paper seems to embrace the logic that regulation hinders innovation, and that is not the case, especially in a high-risk context. The lack of regulation will make risk more concrete. It will lead to breaches of human rights, accountability and responsibility, which will in turn prompt scandals and lower even further the trust the public has in AI in this country. This is very problematic.

It is problematic when we adopt this approach to areas such as autonomous weapons systems in defence, because we miss the opportunity to have central discussions about the regulation of key elements of the use of weapons in our society. This is made further problematic by the idea, which percolates in the paper, of using non-statutory measures. These are important when we talk about actions over and above legal compliance, but there are some cases, such as weapons, where we do not have legal compliance set yet. We need to go back and think about centralised or governmental level discussion. Following the White Paper when it comes to defence might lead us to too little regulation in this area.

The Chair: Thank you very much. We have run out of time. I would like to thank our witnesses very much indeed for spending time with us this morning. Professoressa, mille grazie per la vostra competenza. Dr Blanchard and Verity Coyle, thank you very much indeed.