HOUSE OF LORDS

# Artificial Intelligence in Weapons Systems Committee

## Corrected oral evidence: Artificial intelligence in weapons systems

Thursday 30 March 2023

10.05 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.

Evidence Session No. 2 Heard in Public Questions 15 - 23

### Witnesses

I: Yasmin Afina, Research Associate, Chatham House; Vincent Boulanin, Director of Governance of Artificial Intelligence Programme, Stockholm International Peace Research Institute; Charles Ovink, Political Affairs Officer, United Nations Office for Disarmament Affairs, United Nations.

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

Yasmin Afina, Vincent Boulanin and Charles Ovink.

**The Chair:** Welcome, everybody, to this sitting of the committee. For our witnesses this morning, just to let you know, the session is being broadcast and you will have the opportunity to check the transcript afterwards for factual accuracy. We have a bit short of an hour and a half in front of us and we have a lot of ground to cover. I would like to dispense with any opening statements. We will go straight into questions and you will have the opportunity to add to your answers as you wish.

Q15 **Lord Clement-Jones:** Good morning. I will start off with a very broad, general question. What do you see as the greatest possible benefits and most pressing concerns associated with autonomous weapons systems, which no doubt we will shorten to AWS throughout our discussion today?

***Charles Ovink:*** I would like to thank the committee very much for having me here as a witness. I will keep my remarks as short as possible to allow as many questions as possible. I represent the Office for Disarmament Affairs, so naturally our focus is more on the concerns presented by military use of AI, including autonomous weapons systems (AWS). Although it has been argued that there could be improvements in accuracy or reductions in collateral damage, this has not yet been demonstrated.

I would instead point to the potential impacts that verge more on the concern side. AI applications in the military domain may lead to international instability. They have the potential to introduce elements of unpredictability at times of international tension. They can lead to actions that are difficult to attribute. Difficulty of attribution is an element that is likely to come up frequently today. This can create risks for misunderstanding and unintended escalation, which I think you can also agree is a serious concern.

AI technologies have the potential to aid decision-makers by allowing faster real-time analysis of systems and data, and providing enhanced situational awareness. However, this presents its own concerns. It may compress the decision-making timeframe and lead to increased tensions, miscommunication and misunderstanding, including, of particular concern to my office, between nuclear weapon states. That introduces another dynamic that we would be significantly concerned about.

All uses of AI in the military contexts present concerns regarding predictability. The black box nature of the technology makes this particularly relevant for AWS. One element I would especially like to underline is the importance in this context of effective testing, evaluation, verification and validation, and of demonstrable reliability. All of these are essential before deployment. Although there is a general acceptance among civil and military responsible development practices that AI should be robust, by which we mean reliable and safe from a technical standpoint, the testing, evaluation, verification and validation necessary

to ensure this point is a particular challenge for AI-related technologies in a military context and especially for AWS. How can it be ensured that any testing carried out sufficiently meets those criteria?

This is because it is not currently clear that there are any proven effective methods for testing non-deterministic—by this I mean less predictable and more adaptable—systems, such as one would hope AWS would be. Runs of the same test may not generate the same results. The differences between the testing environment and the deployment environment can cause unpredictable outcomes. Once deployed, changes in the data or the algorithms themselves can lead to changes in behaviour, presenting still further challenges.

There is also the broader issue that some AI-related technologies can be relatively easily acquired by non-state actors and so maybe be employed in weapons or non-weapon-related functions. Such actors may not have the interest, or even the capacity if they had the interest, to employ AI-related technologies in a responsible manner. The responsible design, development and deployment of AI should include protections for civilians.

**Lord Clement-Jones:** I will come back later to tease out whether it is essentially the autonomy or the intelligence that is the issue. That was quite a catalogue, Yasmin. Do you agree with that?

***Yasmin Afina:*** Yes. I would echo a lot of the points that Charlie made. Before we get into the opportunities and risks, we need to broaden the scope a little bit beyond autonomous weapons systems. Beyond autonomy as a function, we also need to look at artificial intelligence, which is in line with the committee's mandate here today. Beyond weapons systems, we need to think about AI-enabled programmes, such as software to enable strategic decision-making and those used across the targeting cycle for intelligence collection and creation, for example.

In terms of the key benefits, there are four I would list, obviously with the caveat that the programme works as intended. First is the increased situational awareness, as Charlie has mentioned. We can have, for example, software combining data gathered live from the battlefield, historical data and success parameters. It will increase not only situational awareness but chances of mission success.

The second is better intelligence collection, again with the caveat that it all works as intended and reliably. Thirdly, there is the question of modelling. You can use AI-enabled software for advanced synthetic environment programmes, enabling advanced war-gaming. In the UK, for example, we have Improbable Defence, which does a lot of work in this field. That will enable better training and planning capabilities.

Fourthly, the main advantage is speed, especially for time-sensitive decision-making processes, especially in the nuclear realm. Then again, you have all the risks associated that Charlie has mentioned.

At the risk of sounding a bit repetitive, there is a question of the risks and the assumption that the AI is working as intended. I would outline four. First are risks pertaining to the overhype, and problematic and misleading assumptions around AI. We talk about artificial intelligence. We have this intelligence, this construct, that is inherently based on a human trait, but is it intelligent?

Secondly, linked to that is the issue of risk perception and how it affects the way we evaluate risks, especially now with the AI arms race dynamic, not only in the security and defence sector but in commercial contexts. We can see here big tech companies, such as Google and Microsoft, deploying AI technologies that are perhaps premature in terms of testing and reliability evaluation, at the expense of robust and thorough safety standards, just for the sake of commercial gain. There is also the issue of compliance with the law and ethics, which I think we will cover a bit more in detail later in the hearing.

Third is the integration of the systems with established command, control and communications systems. Coupled with that, there is the issue of technical literacy of the users of such advanced systems.

Fourth are the more general technical issues that may affect the reliability of the system, for example algorithmic bias, black box, the quality of training and testing datasets. From this issue, naturally you will think of ethical, legal and policy issues that we will cover later on in this hearing.

**Lord Clement-Jones:** One of the interesting areas that you covered there was the perception point. Apart from the reality, there is this issue about the perception of AWS, which is added to all the concerns that Charlie mentioned. Vincent, do you agree broadly with those benefits and risks?

***Vincent Boulanin:*** Thank you for the opportunity to give evidence. I would generally totally agree with what Charlie and Yasmin said. I took the question very literally, so I will focus mainly on the specific case of autonomous weapons, although I agree that it is also important to consider the military use of AI in general.

I would like to start with a brief caveat. The answer to your original question about benefit and risk depends partly on how one defines autonomous weapons. As you may know, there are a lot of definitions out there. The working definition that we use at SIPRI recognises that what makes AWS so distinct and unique—

**Lord Clement-Jones:** Sorry to interrupt. We are definitely going to come to that after this set of questions on the risk and benefit. That is almost the next set of questions.

***Vincent Boulanin:*** I just want to flag one element that is critical for that question. This is a weapon that, once activated, can select and apply force without human intervention. All the benefits and risks, in a way,

stem from that defining characteristic: that at least part of the targeting process is done by the machine.

In terms of benefits, when you talk to military planners they will typically highlight five. First, it would allow for greater speed. You could execute targeting tasks much more quickly than if it was a human-operated weapon. It is potentially useful for very time-critical situations, such as air defence or air-to-air combat, where human reaction time might be too slow. That is one.

Secondly, it allows greater reach. You can deploy an AWS in an operational theatre that is beyond reach, either because it is communication denied, so you cannot go there with remote-controlled weapons, or because it is too dangerous for crewed-fighter aircraft, for instance.

Thirdly, it can provide you with greater persistence. The classic argument is that AWS performance is not altered by things such as fatigue, boredom, hunger or fear. The performance would remain the same over time, which makes them very good for missions that are very dull, dirty or dangerous, such as when you have to defend a specific perimeter over a long time or track targets over weeks or months.

Fourth is greater precision. There is the idea that, potentially, autonomous weapons might be more effective and discriminate than if you are using artillery shells that are unguided. Potentially, they can aim more precisely and allow you to reduce the payload. That could also reduce the risk of collateral damage.

Fifthly, it could allow for greater co-ordination and mass on the battlefield. There is this idea that autonomous weapons can operate like a football team. That they could coordinate in much more co-ordinated way than if they were all human-operated. Arguably, that would allow militaries to reintroduce mass on the battlefield, with large numbers of small autonomous weapons overwhelming air defences, for instance.

These are, typically, the benefits highlighted by military experts.

In terms of concerns, my colleagues already covered them. I would pack them in two categories. First are the humanitarian concerns and the fear that AWS, depending on how they are designed and used, could expose civilians and people protected under IHL to greater risk. There, the fear is that the systems, for instance, would not be reliable enough, or would fail in a way that would harm civilians. The system might misidentify a civilian as a lawful target, for instance, or not be able to recognise people who are hors de combat, who are protected under IHL.

There are also people who argue from an ethical perspective – although this perspective is disputed – that it would be morally wrong to have an autonomous weapon to identify military personnel on the battlefield. On the basis that it would be a violation of the combatants' right to dignity. That point is highly contested in the policy discussion.

That is for the humanitarian concerns.

The second package relates to strategic concerns. These have already been covered quite a lot by Yasmin and Charlie. I would perhaps add the idea that it could lower the threshold of armed conflict. Some states might be incentivised to perhaps conduct operations that could lead to an armed conflict because they feel like, since it is a robotic system, attribution would be harder. I would point out here that it is not an AWS-specific problem. It is basically a problem with robots in general.

One of the other two key things is the impact on the speed on warfare. Charlie already covered that point. That is the downside of having something that allows you a faster reaction time. The second is the risk of use by non-state actors. The idea that low-tech autonomous weapons could be developed by a terrorist group or people who are just putting together technologies from the commercial sector clearly needs to be considered.

Q16 **Lord Hamilton of Epsom:** I want to bring back the whole issue of Elon Musk, yesterday I think, saying that he thought AI should be paused. I would like your views on the viability of pausing the development of AI. How would you get international agreement that had any viability in it at all?

**The Chair:** That is not an easy one to volunteer to answer.

***Charles Ovink:*** It is not an easy one to answer. The call from Elon Musk and some of the other co-signatories is specifically relating to functions around ChatGPT and similar systems. Possibly the best answer I can give here is one that is almost an advertisement for the UN and its general cause in this area.

This underlines a few issues. One is the dual-use nature of the technology. Elon Musk is not necessarily calling for a pause in military development of AI. He is calling for a pause that relates to developments led by the private sector and general development of AI, but the dual-use nature of the technology means that these developments have potential military impacts. They also obviously have civilian and economic impacts.

The key issue from a UN perspective in making any kind of pause or international regulation viable is that it has to be universal. The private sector origins and relatively low bars to acquisition of AI applications mean that there is no real obstacle for anyone to work on the development of AI or for anyone to purchase the applications that come out of the development of AI. That means that the only way to ensure effective governance is through arrangements that are universally applicable and developed through an inclusive process.

Your question sort of underlines that. Even if you were able to get complete agreement from every state, it would need to be a process that meaningfully engaged the private sector, civil society and academia, a huge range of stakeholders, to have any meaningful impact. If you take that to the narrower AWS context, we are looking at what kind of

international regulation is possible, who would have to be consulted and who would actually have to be party to some kind of agreement there.

I will end my remarks by pointing quickly to the issue flagged also by Vincent, which is this issue of non-state actors. You would have to have a situation in which non-state actors' accessibility or ability to use these kinds of technologies were also meaningfully restricted. That is where the dual-use element becomes particularly important. In general, the concern for non-state actors is diversion and misuse. They are using civilian technologies, not necessarily technologies designed with military intent.

**The Chair:** Vincent and Yasmin, are you broadly in agreement with that?

***Yasmin Afina:*** Yes, I would agree with the points that Charlie made. The letter that came out yesterday was really timely in the light of our discussions today. I have a few concerns relating to the letter. It is commendable that they have set up these really ambitious objectives for what they are trying to achieve within the six-month moratorium that they are calling for. They are asking for safety protocols and standards, and a rethink on the risks associated with AI technologies.

Six months is definitely not enough. Ethicists, lawyers, civil society, academics and even parliamentarians have been thinking about this for years now and we have not come up with a solution. How do you expect to come up with robust safety protocols and standards within six months when for years people who have dedicated their professional life and time to it have not come up with a solution?

What is interesting from the letter and one lesson that we could learn is the value of slow AI compared to the arms race dynamic that we are seeing at the moment. I am working on AI and I have a hard time keeping up with the news. I see ChatGPT, GPT-4 and Bard being developed in the large language models realm and I cannot keep up. The value of slow AI is that it would allow us more robust, thorough, testing, evaluation, verification and validation processes that would enable the incorporation of ethical and legal standards within the development of the systems.

Something that would also be valuable from this slower approach to developing and deploying AI is that it would allow space for an inclusive and multistakeholder dialogue that would inform development of AI-enabled technologies. In the end, those technologies are really powerful. When you think about AI, we talk about non-state actors. We think about anyone who can do AI research.

Yes, of course, anyone can do AI research and work on their computer, but, at the same time, if you want to run something that is highly advanced, you need high computing power and hardware. You need a lot of resource, whether financial or human, and not everyone has it. Only a handful of companies have access to these kinds of facilities. The value of slow AI is that we can also rethink the relationship we have vis-à-vis

those companies that have the power to conduct this powerful AI research.

**The Chair:** Is that not precisely the problem? Slow AI sounds like a lovely opportunity. We can all look around and assess the situation as it develops, but it is not in the interests of the people who are developing AI, and putting huge resources behind it, to slow it up. Slow AI is pie in the sky, is it not?

***Yasmin Afina:*** Of course, slow AI might not be in their commercial interest. If we think of ethical and legal requirements, it is also not in their interest, for reputational and liability purposes, to develop something that might, further down the line, have such huge risk implications that they might be held criminally liable because, for example, they did not take into account the due diligence measures they had to go through for deploying systems.

**The Chair:** If you can inject that into the bottom line, that is what commercial developers are keen on. If the possibility of criminal sanctions or reputational damage is something that they can see an economic implication for, that might be a lot more persuasive. I do not see the two joining up, somehow.

***Yasmin Afina:*** That is where regulation could step in. We have these kinds of multistakeholder discussions, for example today in this committee, but also elsewhere. If we have a think more closely with actors from the private sector, the law-making community, the policy-making community and the civil society organisation side, there would be really big value in instilling this culture of responsible AI into the minds of those developing those powerful AI technologies.

**The Chair:** Vincent, do you have anything to add?

***Vincent Boulanin:*** Yes. I will try to be brief, since my colleague already covered quite a lot. I will simply add three points. First, AI as a category is not really helpful. It is so broad and vague that it is difficult to define the contours of that sector. There are so many technologies that may fall under that definition, so it is always important to be specific about the type of application or end use that we are concerned with. As my colleagues mentioned, the idea that was discussed in the open letter was about these large language models, such as ChatGPT, that have a very specific application.

Second is the difference between innovation in the civilian and military sectors. Coming back to this idea of slow AI, the military sector is, in a way, a slow sector, to the extent that there are legal mechanisms that slow down the adoption or require that, if a state wants to acquire AI for high-level military purposes, for instance, it needs to go through a legal review process, where it will be ensured that the system is not prohibited under international law and so on. There are, in a way, safety mechanisms baked into the acquisition process that minimise the risk.

There are no international regulations on the civilian side. It is more the national or regional level, as we see now with the discussion in the EU about the EU Artificial Intelligence Act. One other difference is that, in the civilian sector, commercial actors have an incentive to come to market very quickly. In the military sector, you need to be very mindful of the possible risk of failure,[1] while, in the civilian sector it is in your advantage deploy the technology quickly, even it is imperfect. That is the challenge, I guess.

The third and final point is about how we deal with that and whether a moratorium can be a solution. It could be to some extent, but there are obviously challenges with that, as Yasmin mentioned. Having a technology-centric approach to regulation might not always be helpful. Rather, there should be a focus on norms around elements that need to be considered for the responsible development and use of these technologies, such as general high-level principles that companies would need to abide by as they are developing and deploying these technologies. That is the more future-proofed way of approaching this challenge.

Rather than saying, "We need to isolate this specific technology and determine whether we can keep working on it". We need to say, "let's first define the high-level principles and then, from there, let's consider before we deploy the technology, whether we meet the criteria we have defined?"

Q17 **Baroness Doocey:** I am interested to know what you consider to be the role of the private sector in driving development of AWS and shaping norms around their development and deployment? How do you think that the Government can encourage responsible innovation by the private sector?

***Vincent Boulanin:*** I will follow on from my previous point in a sense. There is a difference between acting in the civilian sector and the military sector. When we talk about the private sector, it is important to be specific about whether we are talking about defence companies or civilian companies that mainly operate on the civilian market.

Typically, weapon technologies are developed by defence companies, so we talk about companies that are pure players in the defence market. They have an established relationship with the defence ministries and are generally aware of the requirements that the Government place on them in terms of safety, the need to develop systems that can be used in compliance with the law and so on.

The challenge is that, now, a lot of innovation is coming from the civilian side. We see a trend where Governments are also considering approaching civilian companies because they are leading innovation in many areas. That is a novel dynamic.

[1] Following the session Vicent Boulanin added that because of the persistent risk of failure you are therefore encouraged not to deploy the technology prematurely.

In a way, we see a model where innovation is being pulled by the civilian sector and now the military sector is trying to find a way to adopt these civilian developments for military purposes. There are a lot of challenges for the military, including how you ensure that what works in the civilian context also works in the military context. Then you end up facing very practical problems. For example, you cannot train the system with the same type of data in the civilian sector and the military sector. For obvious security reasons, you need a different type of data. You need to be sure that the data is properly selected and so on. The safety and security standards are much higher and that is a practical hurdle to importing the technology in the military sector.

That being said, it also depends on the actors. States that are responsible will do what it takes to ensure that, when they are importing technologies from the civilian sector, it is done in a responsible way, so they have properly tested and modified it such that it can be used lawfully and safely. It is certified for use, the user will be trained and so on.

There is also the risk that some actors are irresponsible and do not see a need to be responsible. Rather, they see that the benefit of using that technology is greater than using it safely. That is a challenge when we talk about non-state actors, or, potentially, states that, for strategic reasons, do not see a need to adopt the technology in a responsible way. A race to the bottom in terms of safety is potentially a challenge.

**Baroness Doocey:** Charlie, do you have anything to add? Could you also tell me whether you feel that the military is putting sufficient resources behind AI and, in particular, AWS?

***Charles Ovink:*** I will start by agreeing with Vincent. He raised a number of very important points there. To deal with the first part of the question, it is also very important to recognise that, most of the leading, cutting-edge development in the AI sector is coming from the private sector or academia. This is where the dual-use nature becomes incredibly important. I hope that we get an opportunity to really dig into the TEVV—training, evaluation, verification and validation—issue. Vincent flagged that very briefly, but it might be deserving of a question area of its own.

It is important to recognise also that the nature of modern development often compartmentalises the engineers who are involved. That can make it very difficult for them to follow the second and third-order consequences of the decisions made in development. The reason I flag that is that, as Vincent said, this is, effectively, a relatively novel dynamic. It means that not only do we have a situation where technologies are being developed in the civilian environment that may then be used in a military context, including AWS, but decisions made by people at relatively junior levels in private sector organisations can have significant second and third-order impacts for the military context.

Some of that is related to the training data in question, but it can also be about selecting and adjusting the parameters involved in AI applications, so we have to be aware of the potential longer term and larger impacts of

those decisions. This is part of why norm generation and the development of norms is incredibly important.

To finish very quickly on that first point on the private sector and the Government's role, this is an area that really needs to be underlined. There are a number of civilian-focused responsible AI efforts that tend to deal with issues such as algorithmic bias. They tend to look at things in a justice context or similar, but they tend to have a primarily civilian focus. What are the civilian sector impacts of those decisions and of algorithmic bias? They do not necessary respond to peace and security challenges, which can be related but can be significantly different or have significantly more serious impacts[2].

The question is also how we expand responsible innovation to cover peace and security. This is an area where government encouragement could be really helpful, as the Government are also the source of the development of so much good practice and norms around responsible development and deploying practices in general. For those of us who use systems such as PRINCE2, which—correct me if I am wrong—I understand was developed by the British Government, these are models that become standard good practice without necessarily having any formal international regulation around them.

To give a very brief answer on your question about military devotion of resources to responsible generation, several states have taken steps. For instance, the United States developed the DoD principles last year. There are tools like the AI Risk Management Framework that were developed by NIST, the standards framework organisation in the US, that are all relevant for both civil and military contexts.

It is very difficult to say what would be sufficient in this kind of case. Certainly, it is something that militaries seem to be paying attention to. It is also important to underline the point that Vincent alluded to about the legal requirements. I am sure we will get to this later. Things such as Article 36 weapon reviews require a certain standard or a certain process to be gone through before something could be legally deployed and used. That is a key area to look at as well.

***Yasmin Afina:*** There are a lot of dimensions to these questions. I would like to echo the points that Charlie and Vincent made on the dual-use nature of technologies. We have to go beyond looking at these classical

[2] Following the session Charlie Ovink added that an additional challenge is the extent to which military efforts at responsible AI are visible to the public, the nature of the processes employed, and how decisions about use are made. A further challenge in this regard is the source of the data, particularly training data, employed by the military, and the non-obvious consequences of elements reflected in that data. Datasets of this type are necessarily much smaller and more limited in scope than those used for civilian purposes. For example, a military training an AI using data from its prior human operations may find the AI inferring patterns that are primarily relevant to those specific conflicts and operations, and inappropriate for more general use.

technologies that you would instinctively think have a duel-use nature, for example facial recognition technologies.

Some time ago—I think it was a few months or even a year ago—there was an organisation, Bellingcat, that found information on US nuclear weapons stationed in Europe through a database from an application on mobile phones for flashcards. It is because staff in military bases were using this app to memorise information without realising that their inputs about these nuclear weapons systems would go public into the database of this app.

We saw recently—I think it was last week—that ChatGPT, the chatbot application, has established plug-ins where you can train the system to databases on the internet. Technically, if we think about ChatGPT being plugged into the database of this flashcards app, you can actually ask ChatGPT, "Can you tell me where the nuclear weapons are in Europe? Can you tell me some specifics about it?" You have a lot of security implications around this kind of blurring between civilian applications of AI and military applications.

On the question of shaping norms, we have to be careful about the language that we are using. If we go from the basis that norms are international law rules, public international law is fundamentally made by and for states. Does the private sector have influence on the shaping of international law? It probably does. That is undeniable even in terms of the standards, non-binding instruments and technological progress that influences the way the law is shaped.

Even beyond the military sector, for example in the US, you have a set of soft law. Even if they are negotiating Bills and Acts that would be binding, they would be more prone to a self-regulatory approach, where companies are left to themselves. They are basically marking their own homework and have their own responsibility to respect the law. In the EU, it is a much more generalist, prescriptive and risk-based approach. Where does the UK stand there? We could discuss that in more detail later on.

Essentially, we need to be careful, because, in the end, public international law and international humanitarian law are state-centric. If we have individual responsibility in the criminal law realm, thinking even about domestic responsibility and liability of these technologies, if and when we shape international norms around AI in weapons systems, we have to take into account state responsibility.

**The Chair:** Sorry to interrupt you, but there is a mismatch appearing here, is there not? Obviously international humanitarian law needs to be enforced and developed on an international basis. There is also a role for individual Governments being able to create a regulatory environment and enforce it. The extent to which they can enforce it of course is an open question, but there is a role there for individual Governments, operating independently.

***Yasmin Afina:*** Of course, yes. Even though you have this sort of mismatch, as you said, states have a responsibility for bodies such as the military acting on their behalf to respect the law. Otherwise they could be held liable for violations of international law.

I have a quick point on the push for responsible innovation. There are two things I would like to point out. First, there is a space here for the Government to encourage an anticipatory approach to governance, so focusing on the upstream stages of the technology life cycle, so that they would be compliant with legal and ethical norms by design. That is one approach to push for responsible development of AI.

Secondly, we need to acknowledge the experimental nature of AI and the risks associated with that. We have to think of incentives and effective approaches to take into account its experimental nature. One analogy that Chatham House is currently exploring for AI is the approach that the US Food and Drug Administration is adopting for products related to food and drug safety. That would be based on the phased and gradual rollout of products. Even after the release of the product in the public, you have constant auditing to look at the direct and side effects of this product so that, if things start to go wrong and you can see that from the experimental release of this product, you can potentially phase it out.

Q18 **Lord Fairfax of Cameron:** In view of the considerations that all three of you have outlined this morning, is it your view that an AI arms race is realistically inevitable?

**The Chair:** Yasmin, you have been sketching out for us a sort of Panglossian approach to this: it will be all right in the end. I am not sure how many supporters you have round the table, but what is your view on an arms race?

***Yasmin Afina:*** The problem is that we have this hype around AI. Coming back to the commercial civilian sector, you have this arms race dynamic and urge to innovate and deploy technologies in order to have a cutting-edge advantage. I do not think that we are spared from this dynamic in the military sphere. For example, in Ukraine there is the deployment of technologies reportedly from Palantir and Clearview that are based on AI.

These are highly experimental technologies. I do not doubt that likeminded states or even adversaries might want to deploy similar technologies in the battlefield as a response to this deployment of AI in the battlefield now. I would not say that it is inevitable. I am quite sceptical as to whether we can restrain ourselves from an arms-race dynamic, but, if we adopt a careful approach, whether through regulatory measures or even soft law and policy documents, there is room to restrain ourselves from being carried over by the hype.

**The Chair:** At the moment you are a cautious pessimist. Is Vincent also a cautious pessimist?

***Vincent Boulanin:*** We looked into that topic a while back for a report. I personally do not like the concept so much. It is very catchy and that is

why people use it, but it is a misnomer to say "AI arms race" in a sense. We tried to introduce the idea that it is an AI capability race. All the global major military powers see the value of AI for future military capabilities, but they understand that the race is at the level of building blocks.

The first is ensuring that you collect and are able to develop the data to train very advanced AI systems. Second is talent and being able to train enough people to actually develop AI applications, be it for civilian or military application. Thirdly, there is also access to the hardware component, so this idea that you need some semiconductors that are designed specifically, for instance, for machine learning.

We see that capability race playing out in, for instance, the discussion about export control. The US and other countries see the need to place restrictions on specific technologies that would limit the ability of China, for instance, to keep up in the capability race around AI. There is this idea of trying to slow down competitors by preventing them having access to the key technologies and so on.

We also see moves around this idea of preventing other countries or actors from accessing data and so on. That is where we see potentially a competition going on around these components for developing advanced AI.

As I flagged before, there is also a prospect of having a race to the bottom. Some actors, because they want to be able to embrace the benefits of AI for military capabilities, will disregard norms around IHL compliance or safety, and decide to deploy AI-based systems that are unsafe or not properly designed. If used in armed conflict, that could expose civilians to harm or increase the risk of escalation. That is the real security concern around that competition. Some actors will adopt irresponsibly just because they feel the urge to keep up.

**The Chair:** We all recognise that "arms race" is a rather lazy shorthand, but that the phenomenon shares a lot of characteristics with arms races as we have come to observe them. Charlie, you are in the business, I would guess, of trying to make us a bit more optimistic about these issues. Do you want to reassure us, as it were?

***Charles Ovink:*** I will be as reassuring as I possibly can be. I would take that perspective. We have had arms races and security dilemma situations before that we have avoided, as a planet, and that we have been able to step back from. I would not call it inevitable in that sense.

I completely agree with the point that Yasmin and Vincent have raised. Getting back to the dual-use nature of the technology and this idea of having potential arms race dynamics in the building of capacity, it is important to underline that, given the nature of the technology, even a capacity that is scrupulously civilian may be perceived by neighbours as a kind of latent military capacity. Even if, to some degree, we are talking

about a focus exclusively on developing a significant domestic civilian AI capacity, it will have an impact.

I would briefly like to flag this issue of, to some degree, a kind of jurisdictional fragmentation. The nature of the companies that we have mentioned before means that not necessarily all those aspects are located within a single jurisdiction, whether we are talking about where the data is collected, where the servers are and those kinds of things.

**The Chair:** Jurisdiction shopping is always going to be an option for some conglomerates.

Q19 **Lord Grocott:** I wanted to ask our witnesses to help the committee by giving us how they would define autonomous weapons systems. We have found a number of attempts at that so far, but the briefer, the better.

***Charles Ovink:*** This will be a difficult one to be brief on, but I can give a quick summary. Especially from the perspective of my office, because we provide substantive support to the group of governmental experts—GGE—in this area there is not an internationally agreed definition. There are debates around the extent to which a definition is useful, and different proposals as to whether an internationally agreed definition is even needed or whether characterisation is better.

One of the general points of convergence seems to be that a definition, characterisation or description should be technology neutral—I think that we have mentioned this briefly before—so that it can adequately cover all future types of weapons or weapons systems. It should describe the functions and capacities of weapons or weapons systems, centring on the elements of autonomy and human control, which I am sure we will get to again soon, that are at the core of what is being discussed in these debates.

The technological or technical definition is not necessarily completely required or useful. A good illustration is that the NPT—the nuclear non-proliferation treaty—as far as I am aware, does not discretely define what a nuclear weapon is. It does not necessarily need to. On the technical definition- those weapons systems that function without human control or oversight and cannot be used in compliance with international humanitarian law should be prohibited and all other types of AWS regulated. This responsibility cannot be transferred to machines. These kinds of elements are useful definitionally.

It is worth recognising that a rigid definition, even at a domestic level, that may later become an obstacle to joining an international instrument, may also not be advisable. I will finish by taking a step back further and noting that definitions of AI and autonomy, as I am sure the committee has found, are also not universally accepted in general. To some degree, AI can be a moving goalpost. Committee members may be aware that, not that long ago, if an AI could reliably beat a human at chess, that would have been accepted as a meaningful definition of AI. In the current environment, that is effectively useless.

***Vincent Boulanin:*** I will try to be very brief since I mentioned this before. The working definition that we use in a way reflects a lot of the working definitions that are there. It recognises that what makes an autonomous weapon unique is that, once activated, it can identify, select and apply force to targets without human intervention. We think that that is the most productive working definition. It is not to say that it is the definition that needs to be used for potential regulations, since definitions have a political component to them. For the legal, ethical or security analyses, these are the important elements.

I can add a few additional elements that are usually associated, in terms of characteristics, with autonomous weapons. It is good to remember that AWS can take all shapes or forms, be equipped with very different types of payloads, and be used in very different types of missions and conflicts. That is one of the challenges in the debate. People think of different types of autonomous weapons. People might think of high-intensity conflicts and drones that will fight at machine speed, or more counterinsurgency situations with drone models that track down some terrorist group somewhere in a cave. These are totally different applications.

Another important characteristic for the analysis is that it is a pre-programmed weapon. What makes autonomous weapons different from other guided munition is that you programme the weapon to attack target types or categories of targets, rather than a specific designated target, so saying "these categories of tanks" rather than "that specific tank in that specific location".

It is also a target activated weapon. When in the environment, once it has recognised the type of target to engage, it will engage automatically. The human does not need to press a button. That being said, for both characteristics, the human may be in a supervisory role and retain the ability to override the system, but they do not need to. The system can implement the targeting cycle automatically. That is to say that the human can be in a supervisory role and they can still be autonomous weapons, in my view.

**Lord Grocott:** So much of this is speculative. I am finding that these kinds of discussions are almost academic. You all have difficulties in defining AWS, or at least that seems to be a pretty consistent problem throughout the whole of these discussions. However you define it, could you please give me an example of an existing weapons system that either meets or comes close to meeting the definition you have? I would like something I can get my hands on.

***Yasmin Afina:*** I will not be able to provide a specific example of an autonomous weapons system or an AI-enabled weapons system, but there is evidence that there are other military applications of AI-enabled systems used in the battlefield. For example, we have reports of technology developed by Clearview—that is the company—being used in Ukraine to identify dead bodies, which, in principle, is in line with international humanitarian law requirements to dispose of the dead. It is

based on facial recognition technology, and that is established as an AI. I would not say it is a weapons system because it is used for other functions.

There are also reports that Palantir is using AI-enabled technologies to help with selecting targets in Ukraine.

**The Chair:** Lord Grocott is after something very specific here. Can you point to a piece of hardware that fits into one category or not? We are well aware of some of the developments you have described, but this seems to be the block in our route to understanding and appreciating exactly what we are talking about.

***Vincent Boulanin:*** There are a few historical examples of automated weapon systems that, according to the definition we use, could also be described as autonomous weapon systems. Once activated, they can select and engage targets without intervention. These are typically defensive weapon systems, and there are different varieties of them.

For instance, you have the Phalanx close-in weapon system, which is a system built on a ship that can be turned to autonomous mode if there is an incoming overwhelming attack. The system will automatically identify incoming threats and try to neutralise them, be they incoming missiles or planes.

There is a version for tanks, called an active protection system, where you also can activate an autonomous mode. If they are incoming threats to the tank, it will try to identify the munitions and neutralise them. These types of defensive systems have been around for a very long time. They can also be used in a supervised semi-autonomous mode, where the human authorises the use of force, but they can be turned to a fully autonomous mode.

The other more offensive type of weapon, the one that is closest to what people are concerned about, is what we call loitering munitions. This is basically a mix between a drone and a missile. The Harpy was developed in Israel in the 1990s, but similar systems have been under development over the last few decades.

The idea is that you send the drone over a specific area and you have it loiter above that area to find the type of target it is supposed to attack. Originally, they were designed for the suppression of enemy air defence. The Harpy, which I mentioned, is supposed to find mobile radars. The problem with mobile radars is that you do not always know where they are. They are switched on; they emit for a limited period; and then they switch off. The loitering munition hovers over an area and, as soon as it detects a radar, it destroys it.

These have been around for a very long time. We can now see the potential development of loitering munitions that can be used against other types of targets such as tanks. There have been research projects that have tried to look at that for a long time.

An interesting anecdote is that, when these systems were developed in the 1990s, they were originally supposed to be fully autonomous. It is called "fire and forget". Once it is sent, the system will just do its thing and you will not be able to remotely control it. Over time, the new versions have been equipped with remote control. The operator keeps the possibility of steering the system and aborting the mission.

It is an interesting development. We do not necessarily see a trajectory towards more autonomy. Militaries still want to keep some control over the weapon, if possible.

Q20 **Lord Houghton of Richmond:** Good morning. Thank you for spending time with us this morning. Can I also ask you to be helpful to the committee here, not cautionary but very positive, in making some recommendations about human involvement? I say that because the committee's problem is not really anything to do with ethics, morality or compliance with IHL. We could probably all sit degree qualifications in them, given the amount of evidence we are taking on those things. Our problem is the challenge of how best to comply while still maximising all the benefits of AI, which we know are out there. Vincent, you mentioned cost, precision and potentially having no human military casualties.

From my perspective, there are probably three routes to compliance, which certainly do not involve the creation of new law. One of them, which you have mentioned, is selective regulatory bans on the basis of legality or because these systems employ a black box technology that is not understandable even by the experts that develop it.

The second route would be to place technical limits on the degree of autonomy on a fielded system so that at some point there was human involvement in the chain.

The third route is any enhancement to what I would call battlefield regulations. After all, the battlefield is highly regulated already in respect of rules of engagement, targeting directives, and command and control protocols.

The question is about translating policy into operational reality. To make this possible within the UK policy guidelines thus far, we have to have appropriate human involvement or meaningful human control in relation to AWS. Could you give your own understanding of those terms and perhaps indicate some beneficial ways in which human involvement might be brought about?

***Vincent Boulanin:*** I will start with the first concept you mentioned, context-appropriate human involvement. For me, this term recognises two things.

For an attack with AWS to be lawful, humans should be involved in determining and helping ensure that the effects of the attack are not prohibited under international law, and that they comply with the rules governing the conduct of hostilities, notably the principles of distinction, proportionality and precautions.

That formulation also recognises that the way and extent to which a human needs to be involved may vary. That is the context element. It depends on contextual variables, notably the characteristics of the system, such as how predictable it might be or how it recognises the target, but also the characteristics of the environment, such as how predictable the environment is or whether civilians or civilian objects are present. To what extent might civilians pop up at some point during the operation? That term is supposed to capture that in practice.

Moving on to the concept of meaningful human control, as you will know, the term was originally coined by a UK-based NGO called Article 36. It has become a very popular term in the policy conversation because it captures the idea that humans should keep agency over the decision to use force and their role should remain meaningful. The point of adding the qualifier "meaningful" is that having humans in a supervisory role might not be enough, notably because of problems we know of such as automation bias.

Humans should exercise some kind of active role, to ensure legal compliance and ethical acceptability while also ensuring the mission is efficient from a military perspective. At the same time, it is important to acknowledge that "meaningful human control" is a contested term in the policy conversation for a number of reasons.

The key reason, if I have understood this correctly, is that some actors are concerned that the word "control" could be interpreted in a very narrow way and be equated with this idea of active control over the weapon at all times. For some of the actors, that form of control is not appropriate, because it is not legally required and, in some contexts, may be counterproductive and lead to more civilian casualties.

I do not know whether you want me to elaborate on how this term could be exercised in practice, but, when it comes to exercising appropriate human involvement, based on the work we have done, the idea is that "involvement" means that the human should try to exercise some form of judgment to determine, in light of the characteristic environment, whether the system can be directed at the military objective and the effect will be compliant with the principles of distinction and proportionality.

Depending on different variables, the user may need to take precautionary measures. For instance, they could decide not to deploy that weapon because the risks are too high. They could try to use a different type of control measure to limit the effect and ensure that it is lawful.

In our view, it is important to see that there are a lot of different possibilities within the notion of control. You can try to work with the weapon parameters in the design phase, for instance, by limiting the type of targets a system can engage or how it recognises a target. At the moment of activation, you can try to limit the scope of operation in terms of time and space, so loitering for only 10 minutes rather than three

hours to reduce the risk of unpredictability, for instance. That is one type of control measure.

The second would be to have control over the environment. You can try to structure the environment to minimise the risk of civilian casualties, for instance. You can issue warnings or put up fences to make clear civilians should not be accessing that area. That could be the case, for instance, if you were in a no man's land area or were using an autonomous weapon to protect a specific parameter such as a base.

The third type of measure—this is the one people are more focused on—is human-machine interaction. In some contexts, it might be needed for a human to remain in a supervisory role and keep the ability to intervene to minimise the risk of unintended engagement or disproportionate effects.

These are the three big categories we identified in our work.

**Lord Houghton of Richmond:** Thank you, that is quite comprehensive. Charlie, do you want to add anything?

***Charles Ovink:*** I also think Vincent has been pretty comprehensive. It is important to touch on the issue of the environment itself. Phalanx, the example Vincent gave before, is a relatively uncontroversial system because it is intended to be used at sea. It is in an environment where there should not really be civilians present.

It is important to recognise that it has not been demonstrated that any algorithm can reliably make the human-like decisions and judgments required to comply with the international law. Many states have argued that no weapons system using AI or autonomy in its critical functions can perform those judgments and conform with international humanitarian law.

From our side, the Secretary-General has been pretty clear. He said that machines with the power and discretion to take lives without human involvement are politically unacceptable, morally repugnant and should be prohibited by international law. He was referring specifically to weapons that could be used to target and attack without any involvement, control or oversight of a human operator.

Article 36 weapons reviews have come up again. This is an area my office wishes to explore further. I would encourage states to share good practices in their weapons review processes in this context, especially to support other states. It may also be useful to think about the implications of using AWS and other forms of military AI if their compliance cannot be satisfactorily assessed. What would be the implications for regulation, if you could not demonstrate that they can be used? In addition to the points Vincent has made, it is important to underline that.

**Lord Houghton of Richmond:** Before I come to Yasmin, can you just be clear on that? You almost give the impression that there can be no human delegation to a fully autonomous system. If that fully autonomous

system has been tested and approved in such a way that it does not rely on black box technology, where constant evaluation has proved that the risk of it not complying with the parameters of international humanitarian law is acceptable, there is effectively then a delegation from human to machine. Why is that not compliant? Why would you say that should be prohibited?

***Charles Ovink:*** Again, as I am coming from the UN side, I would be deferring to member states as to what the end decisions are on a total prohibition or what any international instruments might be. The key issue is that that is a very big "if": if it is possible to demonstrate all of those things. Testing, evaluation, verification and validation have come up for a good reason.

The environment for testing civilian AI can be incredibly similar to the environment for the deployment of civilian AI. For an autonomous weapon system, one has to think, "What could the testing environment possibly be like that would make, in this case, your committee satisfied that the system would definitely behave in that specific way when used in the environment of deployment?"

When you talk about delegation, the issue is also accountability. If those things were able to be demonstrated, so you had a system with no black box element that was completely explainable and we could understand why the decisions were made, there would still need to be an element of human accountability. That is the part I would wish to underline.

**Lord Houghton of Richmond:** I am saying that there still is human accountability, because a human—ultimately a politician, a Minister—has given a directive that he is content to delegate to this particular piece of autonomous machinery in certain circumstances such that it can act on its own predetermined algorithms or whatever, so long as they are not a black box that nobody can understand.

***Charles Ovink:*** In that case, the person making that decision would be legally responsible for the consequences of the decision.

**Lord Houghton of Richmond:** Yes, as Ministers are all the time for the authorisation of lethal force. It is one of their duties.

***Charles Ovink:*** There is also a question, which will probably come up in other sessions you have, about how accountable that person can be to the extent that they understand or reasonably could be expected to understand the deviations the system might make.

To go back to what I was saying before, the main issue there is that is a very big "if". How would we demonstrate all of those things? How do we ensure that the politician in this case, or the military officer in other cases, is reasonably informed and could reasonably expect the system to behave in the way it does, but then, if it does not behave in the intended function, it is reasonable to hold them accountable.

**Lord Houghton of Richmond:** I am tempted to observe that in this

respect sometimes the human being is probably as vulnerable to not getting it right as the machine. Anyway, that is a different set of answers.

***Yasmin Afina:*** There will be situations where human involvement, oversight, approval, control or whatever term we are using would be more desirable or even necessary for compliance with international law against the technical capabilities and limitations of the technology we are talking about, such as targeting and engagement in areas with dense civilian populations or where there are heightened risks of disproportionate civilian casualties. On the flipside of the coin, as you rightly pointed out, there are some situations that may require such speed that the cognitive limitations of human beings may constitute more of a liability for legal compliance.

Accountability is not necessarily the question, in the sense that there will always be a human accountable. That is for sure. It is more a question of choosing the appropriate contexts in which there would be more benefits than risks of deploying AI that may have certain levels of human control or not. In the end, it is a question of risk assessment and a calculation of risk versus opportunity, whether it is related to mission success, legal compliance or compliance with morals, ethics and policy.

Charlie mentioned the reasonable expectation to understand how the technology works and the ability to operate human control when necessary. One way to do that is at the design stage of the technology. When we conceptualise an AI system, we have to ask whether it is really worth delegating this task to an AI or whether it is worth keeping it within the realm of human intervention. If it is worth using AI, we have to work around the technical parameters of the technology and the interface that is being used.

In the end, the end users might not have a computer science background. They might not have been involved in the coding, design, testing and evaluation stages of the technology. They are just being handed the technology. One way to ensure they can exercise human intervention or control in the appropriate context is to make sure these interfaces are user-friendly and in a way that will enable them to be aligned with these legal norms.

The other day I was talking with someone from a technology company specifically developing software to help with targeting. One of the solutions they found was to have a pop-up that comes on to the screen and freezes all decision-making until a human operator says, "Yes, this is a lawful target", or that it is not. These kinds of technical solutions and interfaces would be helpful in terms of operationalising the exercise of human control when appropriate.

**Lord Grocott:** This issue keeps arising as to whether the technology does what it is intended to do. That is not specific to AI, is it? A bayonet either works or does not work. If it bends when you are using it, it is not much use. Is that not that just an assumption when you look at any weapon or any weapon system?

***Yasmin Afina:*** Yes, of course. If you look at offensive cyber capabilities, for example, you are going with the assumption that this capability will work as intended and will not go beyond the targeted systems it is intended to be offensive against. It might also be related to missile systems, missile technologies or other kinds of technology. That is why AI is about novel questions but also old questions related to, for example, technical reliability.

Q21 **The Lord Bishop of Coventry:** Thank you very much to our witnesses. It is always fascinating to follow Lord Houghton. My questions are on the same theme but of a much more amateur sort. I am always fascinated to hear Lord Houghton talk about commander responsibility and battlefield regulation.

What I am finding quite difficult to get my mind around with AWS is that, if I understand it correctly, an element of the definition is adaptability to the environment and, as Yasmin said earlier, taking on board battlefield data. Therefore, I am finding it very difficult to see how meaningful human involvement can be integrated into an evolving weapon system. The commander—I may be mistaken, but I want to pick up on your comment a moment ago, Yasmin—may not always have the technical expertise that others would have at different stages of the development. How would that commander, in an evolving situation, with an evolving weapon, be able to calibrate the right human involvement? Would it change as the weapon changes?

I want to ask a bigger supplementary. One thing that I have certainly been very interested in has been the idea that we are not there yet, as it were. It has been a regular theme. Vincent or Charlie, one of you referred to this a moment ago. The weaponry simply is not at a point at which it can be trusted to meet the requirements of international law. I am also very conscious, as Yasmin said, of the speed at which it is developing. You have said you cannot keep up with it. My goodness, none of us can.

I am wondering how we handle that. How should policymakers, lawmakers and even the military think forward to those situations that may not be as far away as we are being assured they will be? I am thinking about the reassuring comments that our Ministry of Defence make about not wanting anything to do with really advanced weapons that really are autonomous and can make their own decisions and so on. We were told that we need not worry about that because it is a long way off. Do we not have to be prepared for the long way off, which might be nearer than we might anticipate.

That is a very big question. The first one, if you are able to focus on that, relates to this whole question of self-evolving weapons and what human responsibility means in that context. The second question also perhaps relates to that. This does relate to the speed of development because these things are changing so quickly. The capacity to self-evolve is increasing all the time. Forgive me for a rather long set of questions.

***Vincent Boulanin:*** You raise a number of very big and interesting questions. The one point that is important to keep in mind is related to the notion of adaptability. Sometimes we confuse the idea that the weapon may adapt to a dynamic situation and variations in the environment and the ability to learn and change its programming.

The first option might be fine, as long as we can predict and have an idea of how the system may or may not adapt to variations in the weather, for example, or change its behaviour if there are civilians present or not. The system could trigger a different type of response. That can be pre-programmed and controlled in a way that means you still have some predictability. You can predict that the system will behave in a specific way if something happens.

If we are talking about systems that keep learning, taking in new data and re-changing their parameters of use, that would be problematic. People would argue that this weapon would be inherently unlawful because you would need to do a new legal review to verify that the learning has not affected the performance in a way that would make the effect indiscriminate by nature.

If there were to be some learning, most likely that would be in this form: the system learns during the mission; after the mission, the information is uploaded; the system is retested and reassessed; and then it can be used again. You would have to go through some form of screening process. Otherwise that would be legally problematic.

I also foresee that commanders would not want to use a system if they do not know how it might behave or if it might learn something that could lead the system to behave in a way that is unpredictable. That could either lead to civilian harm or a situation where it would be operationally counterproductive. These would be my two main points.

I want to add one last point about the IHL question because it is an important element. There is a growing understanding that we should not anthropomorphise autonomous weapons. Weapons do not comply with the law. The law bites on states and their human agents. The responsibility to comply with IHL is with the states and armed forces that decide to deploy weapons. The evaluations demanded by the principles of distinction and proportionality, for instance, cannot be fully automated in the sense that humans, as the agents of the states, ultimately need to make these decisions. That is one interpretation that is fairly shared in the committee of legal advisers.

**The Lord Bishop of Coventry:** Thank you very much, Vincent. That is very helpful. I know Charlie has spoken about the whole concept of weapons reviews. Am I right in thinking that a weapon may adapt in the usage of the moment that it becomes something different? Therefore, it becomes problematic from a human involvement point of view. Charlie, do you have any further thoughts?

***Charles Ovink:*** I would second the points Vincent raised. In relation to your follow-up question, yes, that is one of the issues at hand. The idea of a non-deterministic system that will adapt based on the surroundings/environment is that it will do so in a way that is presumably intended to be useful in the deployment of the system but also adds another element of unpredictability. This is why Vincent brings up the point that it would presumably have to be tested and evaluated at each steps.

Aside from issues of human dignity, which Yasmin also brought up, and the extent to which it is reasonable, legal or appropriate to assign some of these tasks or roles to a machine, there is also the issue of adversarial action[3] as it comes into predictability. We have been talking about a lot of these things generally, in the sense of a system behaving as we expect it to with no other external issue, but we also need to be aware that there will be vulnerabilities presented by these systems that make them vulnerable to adversarial action, which itself can bring in new elements of unpredictability.

That also ties in to the point that was raised about a bayonet and the ways it might fail. There is a relatively narrow range of ways a bayonet might fail. A person deploying the bayonet could reasonably be expected to anticipate and react to those. You may be familiar with some of these examples with self-driving vehicles and so on. Very minor changes in the pattern of stickers that are put on a stop sign, for example, can mean it is interpreted completely differently by the autonomous vehicle. That is without any intentional adversarial action. You also then need to consider adversarial action in this context.

In addition to the other black box and "big if" things we have covered before, you also need to remember that non-state actors are an issue in this context. From a technological perspective, non-state actors are probably already in a position to use these kinds of systems to meet the requirements that have been brought up: to select a target and to engage with a target.

The question on the state side is the extent to which that is possible while meeting the legal requirements and doing so in a way that meets the requirements this committee is discussing. That is the point I was raising before from the UN side, which has not been demonstrated.

***Yasmin Afina:*** The first question you raised was this fascinating question about how to maintain meaningful human control in the light of the adaptable nature of AI and how it combines a lot of battlefield data. That trickles back to the question of interface.

At the risk of oversimplifying it, when you look at navigational apps on your phone, such as Google Maps, you say, "I want to go from point A to

[3] Following the session Charlie Ovink added that Adversarial action in this context includes inputs specifically designed to look "normal" to humans but that cause misclassifications by a machine learning model.

point B". The software will combine a lot of live data—traffic, bus timetables, incidents, Tube strikes or whatever—and then provide you with a proposed itinerary. You can play around with the proposed itinerary based on a set of parameters you have control over.

If we take this analogy and apply it to military applications, having appropriate parameters will enable users to filter out data that they might not necessarily need to know on the first basis. If they have some control over the parameters and they are able to play around with the outcome, that would be a good solution.

Echoing Vincent's point on keeping up with tech advances, Article 36 legal reviews are useful for assessing the compliance of systems with the law. That does not mean they will always be compliant once you have given the green light because of their ever-evolving nature.

That is why, again echoing Vincent's point, it would be useful to establish monitoring and auditing requirements on a constant basis. Whether the auditing processes would be independently conducted or not is another question. You have the question of industrial secrecy and the extent to which independent parties can play around with a system because of its inherent coding and, again, because of technical and secrecy considerations. You do not want people to mark their own homework as well. That is a consideration.

Q22 **Lord Browne of Ladyton:** The questions I want to ask were anticipated by the exchange between you, Charlie, and Lord Houghton. They are about the sufficiency of international humanitarian law to govern—I will very carefully watch my language, Vincent—the use of AWS, rather than the AWS themselves, and whether we need other international agreements or accords.

Let me just make this comment. I inferred from the UN Secretary-General's remarks that he was talking about technologies that he knew were likely to be developed, rather than just talking in a vacuum, when he was talking about weapon systems that were publicly unacceptable and morally repugnant. This suggests to me that he, as the leader of the United Nations, believes this issue of sufficiency is being challenged. Perhaps, Charlie, you could expand upon what you said earlier to start us off on this.

***Charles Ovink:*** That interpretation is indeed broadly correct. This goes back to the technological question I raised before. From a purely technological perspective, facial recognition and other AI technologies do make viable a system that can select a target and then engage it, as in they make it technologically viable. All of the other issues presented today, whether these are more on the ethical side or are even just practical questions around consistency, predictability, training and son, call that viability into question.

The concern from the Secretary-General is not that these are far-off technologies or this is a far-off potential application or that this is something that is being considered purely as a hypothetical in a vacuum,

but that we are essentially at this juncture. Some weapons systems that already exist could potentially be argued to exist within this space, whether they are used in that sense or not. In all cases, we are certainly within the technological threshold of being able to deploy weapon systems that can do some or all of these things.

The question then is the relevance of IHL, which obviously applies, and how we can ensure compliance both in this sense and in relation to some of the questions that have come up about predictability.

There is a growing understanding that States should prohibit autonomous weapons systems that cannot be used in compliance with international humanitarian law. Any other type of weapon that incorporates autonomy into the functions I have mentioned, selecting and attacking targets, should be strictly regulated. These regulations could include some of the things that were brought up before, such as limits on targets and locations in which such weapons can be used. Whether that comes in the form of an international instrument or in some other form is still an open question, but the important point is that international humanitarian law continues to apply fully to all weapons systems, including the potential development and use of autonomous weapon systems.

It was also flagged earlier today that such weapons and weapon systems must be able to be traced back to the human element to ensure accountability and dignity and to address some of the other questions that come up.

Such as around  the implications of the Martens clause. Regardless of the state of the technology, under international humanitarian law, even in situations without an explicit rule, civilians and combatants remain under the protection of principles derived from public conscience.

The question is not so much whether international humanitarian law is sufficient to cover AWS. We have clearly covered lots of examples of the relevance and importance of IHL and international law more generally. It is about how that is operationalised and what that means for the context of this committee.

One of the questions I raised before in the context of Article 36 weapons reviews was thinking about what it would mean if compliance cannot be satisfactorily assessed. What does that mean for the development of any of these kinds of systems? What does that mean for the development of any system that conceivably has some of these technologies or capacities? I would probably just want to underline that element of it, and then I am happy to pass on.

**Lord Browne of Ladyton:** Yasmin, I want to expand this question, and I am encouraged to do so by one of your first answers. Beyond the actual weapons system that is deployed for its lethality, should we expand this to include lethality-enabling systems in weapons? What is your view as to whether international humanitarian law, as presently applied, is sufficient to govern the use of those?

***Yasmin Afina:*** Yes, we should definitely expand our thinking beyond weapon systems to all systems that are lethality-enabling.

For example, if you look to the other side of the pond in the US, there was this revelation about the National Security Agency having a programme that collects data from mobile phones according to selected parameters. Based on the data from people's mobile phones—the location, the call logs, et cetera—they could identify an al-Qaeda courier. It is just based on the routes the mobile phones are taking. Their itinerary would suggest that someone was a courier for al-Qaeda. When they deployed the system, they mistakenly identified an Al Jazeera journalist as an al-Qaeda courier.

That kind of system is not a weapon system, but it is lethality-enabling. If you were to engage the target, the journalist, that would absolutely be against international humanitarian law considerations. I would definitely agree that we should be covering these kinds of technologies.

Is IHL sufficient? The law is there. It was designed to take into account all forms of warfare, even if its nature changes. Even futuristic scenarios would be covered by existing rules. I would be careful about saying that the law is not sufficient. That would imply that there are gaps in the law that people could be exploiting for their own gain. That would go against the rule of law. The rules are versatile because they are general. There is a lot to unpack, and there is a lot of clarification that would be needed and would be useful.

We also need to think about how it would be done. It could be done through international processes, a treaty or soft law. We have to take into account that it is incredibly difficult to come up with an international treaty, especially now, in this geopolitical context. We can see in the nuclear sphere how contentious it was to negotiate the nuclear test-ban treaty, for example. If we take it holistically and take into account hard law instruments, soft law instruments and other processes that would enable us to clarify how the law would apply and be operationalised in relation to the deployment of AI-enabled systems, that would be a really useful first step.

**Lord Browne of Ladyton:** I just have one final point. I will ask you this question, Vincent, although Charlie was the one who encouraged us to look at Article 36 as a key area. I will ask you to engage with this question. One way of testing whether the existing rules and regulations we have are sufficient is to look at Article 36. Are you aware of any existing weapons review of an autonomous system that assessed that system as being compliant with international humanitarian law? Secondly, do all those jurisdictions that carry out these Article 36 assessments publish them? Is the public entitled to know what these reviews conclude?

***Vincent Boulanin:*** Thank you for the question. SIPRI has done quite a lot of work on Article 36 in the past years, especially in the context of emerging technologies and autonomous weapons.

The first element, to answer your question, is that there is no requirement by law for states to disclose the outcome of a legal review. It is not that they cannot do it, but most states do not disclose that information. To my knowledge, I cannot report any recent case of a weapon review of an autonomous weapon.

I have heard that some weapons have been reviewed, notably in the US, but I cannot provide any evidence. The only thing I can say is that some of the automated weapons we see today, such as the Phalanx systems or the Harpy that I mentioned earlier, have most likely been through these processes before. That is one thing: we know there are some weapons that share some characteristics with autonomous weapons that have been deemed lawful in the past by these processes.

One element that is important to note with regard to the value of that process is that it is one critical mechanism to ensure that a weapon can be used in compliance, but it is not the only one. I want to stress that the actual legal purpose of Article 36 is to determine whether the employment of a weapon system would, in some or all circumstances, be prohibited by the additional protocol and any other rule of international law.

That is the first thing states will try to do, as they are considering the development or adoption of new weapons. It might be something that is already prohibited, whether by a specific weapon treaty, such as the Biological Weapons Convention, the Chemicals Weapon Convention or the blinding lasers treaty, or any other general rule of international law. If it is not prohibited, it will pass the review, but most likely the reviewer will issue some recommendations on the elements to be considered for the lawful use of the weapon. These recommendations will be really helpful for designing the rule of procedure and the rule of engagement specific to that type of weapon.

Here I want to note another mechanism that we do not talk so much about, but it flows from rule 82 about the provision of legal advice. IHL also demands the provision of legal advice where necessary. That means having legal advisers at the side of commanders to help them assess whether the use of a specific weapon in a specific context will be lawful.

In the context of autonomous weapons, that is highly important and could be deemed necessary. The commander would be supported by legal advisers to assess whether, at the moment of activation, he can make the assessment that the effect could be lawful or at least not prohibited.

Q23 **Lord Fairfax of Cameron:** May I just ask one very quick question to Vincent arising out of something you said earlier? You mentioned Phalanx, and you drew a distinction between offensive and defensive weapons. Is that distinction in fact irrelevant? It is very difficult to work out where the dividing line between an offensive and a defensive weapon is.

***Vincent Boulanin:*** That is a very good question. I would agree with you that the distinction is not very useful, especially if we are getting into the territory of policy-making and rule-making. It is useful for categorising some of the weapons that are in use today, but defensive weapons can also be using an offensive way. How do you categorise them? The division between defensive and offensive is not very clear-cut. I agree with you on that point.

**Lord Fairfax of Cameron:** What one recommendation would you make to the UK Government in this area, if you could make it?

**The Chair:** Lord Fairfax may be looking for a one-sentence answer.

***Vincent Boulanin:*** I will keep it to one sentence. The UK has been very active in the conversation in Geneva on autonomous weapons, especially with regard to pushing for more in-depth discussion on IHL and how IHL applies to the development and use of these weapons. That effort should be commended.

My recommendation is that the UK should keep supporting this type of effort, ensuring that, at a substantial level, the discussion is moving forward and trying to provide greater clarity as to how the rules of IHL need to be interpreted and applied in the context of autonomous weapons as there are still elements that, in some respects, can be open to very different interpretations. It is a very useful exercise.

***Charles Ovink:*** I am certainly lucky that I can take that sentence for granted. I would stress the importance of responsibility throughout the AI lifecycle, including in the civilian space. This development of norms and best practice is something the UK has been traditionally quite good at. I mentioned PRINCE2 and other programme standards that have become international, so there is that background.

Given the nature of the technology, we would emphasise the dual-use issues, embedding peace and security concerns within normal risk management, mitigation and so on in normal AI development practices in civilian and defence situations. That would be something I would heavily recommend and encourage the UK Government to support, especially as this would be building on good practice the UK Government have already led on.

***Yasmin Afina:*** I would suggest being a careful leader in the responsible development of AI. Focusing on the upstream stages in the technology lifecycle to promote compliance by design is one recommendation I would stress. Of course, we should not discount the importance of the post-deployment stages across the lifecycle of the technology, but there is a lot of room for the UK to instil this culture of safety and compliance by design internationally.

It can go a long way, especially in the AI field, where unfortunately legal and ethical considerations are more of an afterthought than something deeply rooted within the research and development culture in the earlier

stages of the technology's lifecycle. That would decrease the risk from the beginning and would provide space for multidisciplinary, multi-stakeholder and inclusive input for the development of these powerful technologies.

**The Chair:** Thank you very much indeed. Vincent, en nom de la commission, merci mille fois pour votre contribution et votre expertise. To all of you, to Charlie and Yasmin as well, thank you very much indeed for the time you have spent with us this morning. We have all found it extremely helpful. As always, we will have various things to follow up as well. Thank you very much indeed.