HOUSE OF LORDS

Artificial Intelligence in Weapons Systems Committee

Corrected oral evidence: Artificial intelligence in weapons systems

Thursday 15 June 2023

4.15 pm

Watch the meeting

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

Evidence Session No. 10 Heard in Public Questions 133 - 138

Witnesses

I: Dr James Johnson, Lecturer in Strategic Studies, University of Aberdeen; Christopher King, Head of Weapons of Mass Destruction Branch, UN Office for Disarmament Affairs.

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

Dr James Johnson and Christopher King.

Q133 **The Chair:** Good afternoon, Christopher King and Dr James Johnson. Thank you very much indeed for coming to join us this afternoon. I am sorry that we are starting a little late. This session, as you know, is being broadcast and you will receive a transcript of the evidence that you give to us so that you can check it for factual accuracy. Could you perhaps very briefly introduce yourselves and the standpoint from which you join us? I do not mean geographically. I mean philosophically, as it were.

Christopher King: At the outset, I would like to say thank you for inviting me to speak today. It is a real pleasure and a privilege for my part. I should also like to say best of luck to the English cricket team for Friday, but I am not going to say that at the outset.

I am the chief of the weapons of mass destruction branch in the United Nations Office for Disarmament Affairs. To put it succinctly, our key goal and our role is to assist the member states of the United Nations in the prevention of use and proliferation, and the eventual elimination of, weapons of mass destruction, with a particular priority on nuclear weapons.

The Chair: When I saw your CV, I thought you might be taking an interest in events at Edgbaston tomorrow, but, on that, we will have to agree to disagree.

Dr James Johnson: Thank you for inviting me today. It is a great pleasure. Good afternoon to the committee. I am geographically calling you from slightly further north, up here in a very sunny Aberdeen. My background is that I am a lecturer in strategic studies at the University of Aberdeen. I am sure you have had a preview of my CV, which I would have cleaned up in advance if I had known. I have a key interest in all issues relating to nuclear non-proliferation: the intersection of emerging technology—especially artificial intelligence and autonomous weapons—with nuclear issues, and especially nuclear command and control issues. I have published quite prolifically on these topics in the last couple of years. My most recent book is *AI and the Bomb*, just published this year with Oxford University Press, which covers all the interesting areas on strategic theory and deterrence, et cetera.

In addition to my role at the University of Aberdeen, I am also a non-resident fellow of the Towards a Third Nuclear Age ERC-funded project hosted by the University of Leicester, and a non-resident fellow with the Project on Nuclear Issues at CSIS in Washington.

Q134 **The Chair:** Can I start with the headline issue? If you combine the availability of AI and nuclear weapons, are we introducing a completely new strand of hazard into world events?

Christopher King: There are a few points that I would like to make at the outset. First, any risks posed with the intersection of artificial

intelligence and nuclear weapons can be avoided through the elimination of nuclear weapons. Indeed, nuclear risks will persist for as long as nuclear weapons do, and the fulfilment of the near universal commitment to the pursuit of nuclear weapons is the best way to ensure that they are not used.

Secondly—I am sure that other witnesses have made this point in the past—"AI" is a very broad term and has multiple subfields. Machine learning is one facet. AI can be used in multiple ways in security and defence—for example, data analysis, computer vision, including image recognition, and guidance.

Thirdly, the use of AI in military applications, especially in conjunction with nuclear weapons, also needs to be seen in the broader context of the convergence of several emerging domains and new weapons technologies that have opened potential new vulnerabilities in nuclear command, control and communications—for example, cyber or space technologies—or compressed decision-making windows, such as the development of long-range weapons with enhanced speed, stealth and accuracy.

Fourthly, it should be noted that the nexus between AI and nuclear weapons contains the same broad concerns about the use of AI in any weapons system: inadequate training, the potential for mistake, poisoned data, the so-called black box problem, unpredictability and compressed decision-making timelines, with the addition of potentially existential consequences.

Fifthly, escalation concerns also relate to the category of weapons system. However, the use of AI in pre-delegation of the launch of nuclear weapons is an extremely dangerous concept that could result in catastrophic outcomes. Although some can argue that pre-delegation might temporarily bolster deterrence, it ultimately increases the risks of accidental or misperceived nuclear use.

Although it might be tempting in an era of increasing geostrategic complexity and technological development to backstop deterrence with so-called dead hands, there is far too much uncertainty and the consequences are far too great. There are multiple political, legal, ethical and moral arguments for the maintenance of human control over nuclear weapons, but perhaps the simplest is the avoidance of extinction. Even when human control over the actual use of nuclear weapons is maintained, it is essential to ensure that the information received by decision-makers is accurate and that they have the longest decision-making window possible.

Finally, equally as worrisome as the potential of AI in weapon systems is the prospect of using AI in attacks against nuclear command, control and communication structures. It is not unforeseeable that AI could be used to spoof, hack or even deepfake early warning systems or other control structures into believing a nuclear strike was under way. The broad diffusion of technology, especially ICT, could enable malicious third

parties or even sophisticated non-state actors to engage in such activity. Concerns about interference in NC3 could create worrying "use it or lose it" scenarios.

Dr James Johnson: I will keep my opening comments brief. On Chris's comments, nuclear weapons and AI-infused weapons within the nuclear domain are not ideal, and the logic and rationale does not suggest that that is the case. In my research and discussions with interlocutors in the US and the UK, there seems to be an inexorable and inevitable trend, driven by several factors, certainly with the harsh geopolitical realities at the moment being a deciding point.

I am sure that we can unpack these further in the discussion, but they include things like first-mover advantage, which was discussed in previous sessions, and the security dilemma between great powers in the current political era, as well as technological determinism. All these aspects will mean that AI will inevitably diffuse within the nuclear domain, albeit at various levels and to various degrees by different adversaries.

On a broader perspective, AI is part of a new and rapidly evolving package of technologies that could enhance and enable a portfolio of weapons such as hypersonics, cyberattacks, counterspace, direct energy weapons and non-strategic nuclear weapons, which has been quite an overlooked part of the discussion. As Chris mentioned, these are complicating escalation dynamics and creating new and novel challenges for strategic stability.

Like others, I also view AI as a powerful force multiplier of existing and new weapon systems, rather than being weapons per se. That distinction is quite important. That is to say that, AI does not exist in a vacuum in this way. Very much like space and cyber domains that came before it, the risk of inadvertent and accidental escalation is heightened in the AI-enabled weapon system due to things such as poorly defined, non-binding or non-existent norms of behaviour, as well as unclear escalation thresholds between nuclear powers, complex cross-domain interactions, and new and increasingly autonomous capabilities that we see.

My view is that AI-enhanced weapon systems operating at very high speeds and increased levels of sophistication and in very compressed decision-making timeframes, as Chris mentioned, will further reduce the scope for deliberation and de-escalation of crisis situations and contribute to a nuclear-first course and other accidents and mishaps.

The use of weapon systems can contribute to escalation conflicts to the point of nuclear war, and there are various thresholds up to that point, albeit that they are essentially pliable and very much in the eye of the beholder. They are perceptions rather than cogent thresholds, for several reasons, including the desire to speed up war, creating new vulnerabilities, especially cyberattacks to NC3 systems, in creating a new range of pre-emptive strikes and very much undermining deterrence and

crisis stability. This is where my research interests lie, and I am sure that we can unpack these in the discussion.

Just to conclude, as a caveat, the harsh geopolitical realities that we are facing at the moment between multipolar nuclear powers certainly makes it very difficult to speak candidly about the AI strategic implications, especially involving the highly classified NC3 systems, which most nuclear powers are now modernising very rapidly, given the importance of military speed and especially countermeasures to these vulnerabilities that have been developed.

The Chair: Thank you. I think you may have covered the area that Lord Fairfax was going to ask you about.

Lord Fairfax of Cameron: I agree. My question may have been covered already.

Q135 **The Lord Bishop of Coventry:** Thank you for your presence today. To what extent are domestic or international regulations required on the use of AI in nuclear command, control and communications? I wondered whether you had any thoughts on the Block Nuclear Launch by Autonomous Artificial Intelligence Act introduced in the US Congress.

Christopher King: Legally binding international agreements are increasingly difficult to achieve in the current geostrategic environment, and there are also justified concerns that any agreement, including on AI and NC3, would not be verifiable or enforceable, for some of the reasons that Dr Johnson alluded to earlier.

However, that does not mean that states possessing nuclear weapons should not be attempt to negotiate such an agreement, because even if such an agreement was not politically binding, if it was agreed by all states possessing nuclear weapons it would still have significant value as a risk-reduction and confidence-building tool, especially if it also prohibited interference with NC3 systems.

The Treaty on the Non-Proliferation of Nuclear Weapons could be a venue in which to have discussions on this topic, potentially as a precursor to including commitment by the nuclear weapon states on this issue in the outcome of the 2026 review conference. In the interim, the issue should be discussed as part of broader risk reduction dialogues and in the context of the nexus between nuclear weapons and emerging technologies and demands that is currently going on in various forums—[*Connection lost.*]

The Chair: I am afraid you have frozen. While we still have you, Dr Johnson, would you like to take up the baton?

Dr James Johnson: Yes, certainly. I think I am still firing on all technical cylinders here. There is no cyber interference that I am aware of.

There are two questions there that we need to unpack. Both are heavily contested, which I guess is the reason for asking these questions in this

forum. Let me address the first one. I certainly agree with all Chris's comments that he covered in his opening statement about the requirement or perhaps the viability of domestic and international relations for the intersection of AI with NC3 or nuclear command and control. Apologies for all acronyms and jargon that are hard to avoid in these kinds of discussions.

The use of AI in NC3 is a complex and potentially hazardous issue that would certainly benefit from careful regulation, deliberation and discussion, not only by the UK but in an international forum. As Chris mentioned, the main challenges are how to ensure getting all heads to the table in this current geopolitical environment and ensuring that the systems that have been developed by various states are reliable, safe and, as the US and the UK, in its recent AI strategy stressed, firmly under political human control.

In terms of potential ideas, I can spin out a couple of suggestions for regulations. These might include ones that stipulate requirements for human supervision and intervention capabilities to prevent unintended inadvertent consequences, or regulations to define who exactly would be responsible in the what I would argue is the very likely event that AI-infused applications make errors or there are technical glitches, whether these are caused by human or machine, or by human-machine interaction, and how responsibility is apportioned in these undesirable outcomes.

AI systems in nuclear command and control also need to be incredibly reliable and safe, so regulations and standards need to be set, including adding redundancies, fail-safes and a robustness against potential accidental failures. There is a new buzzword going around now that sounds quite nice: "graceful degradation". It is the Pollyanna-ish desire or the ability for an AI system to maintain reasonable performance and functionality, even when it encounters novel inputs and situations that you would expect to find in a nuclear crisis situation.

Other things include ensuring transparency and accountability and, more practically, regulations that could specifically ban the use of AI for autonomous launching of nuclear weapons, which has been experimented and discussed by several states, and attacking NC3 systems or satellites that inform or feed into NC3 systems with things such as cyber weapons.

Moving swiftly on to the second question about my thoughts on what is, it is important to stress, a proposed Bill by the US Congress, this very much codifies the 2022 nuclear posture review, which has an existing ban on the use of AI for autonomously launched nuclear weapons and, very much like the UK, stresses the need for humans—the President—to be in the loop in initiating and terminating nuclear weapons. Essentially, it ties federal hands on releasing funds for the use of any launching of nuclear weapons by an automated system without "meaningful human control". It also serves as an opportunity for the sponsors of the Bill to emphasise their efforts towards nuclear non-proliferation. They also have an agenda

and it very much complements their recent efforts to restrict the US President's power to unilaterally declare nuclear war.

In terms of the supporters and the evidence that they put forward, the prima facie evidence for these arguments is quite axiomatic. You would not really argue with the main reasons for the Bill—things like preserving human judgment, given that the decisions needed to launch nuclear weapons require judgment, discretion and accountability, which are things that AI, in its current state of development, clearly lacks. It codifies, makes clear or preserves legal and moral responsibility very much within the human purview.

It also argues that having a dead hand or an automated nuclear launch process would fail to abide by international human rights law. Things like distinction, proportionality and humanity would be required, and it falls at these hurdles. In terms of things like reducing the risk of escalation and arms racing, you would imagine that, if the US had a declared policy to develop automated systems, its adversaries would follow suit pretty swiftly.

In terms of the arguments against the Act, there are three threads. I will not go into too much detail. One is that it weakens US nuclear deterrence. The second is that it reduces the President's time to respond to threats. Counterintuitively—ironically, this one is quite persuasive—it might undermine strategic stability.

In terms of weakening US deterrence, the opponents of the Bill argue that it does not enhance deterrence, which is the main focus of the nuclear posture review by the US. Rather, it seeks to frustrate these efforts across the nuclear enterprise to speed up the processing of data because of an exaggerated fear that intelligent machines would somehow behave like Skynet and wipe out humanity.

The second point is that things like machine-learning algorithms can speed up decision-making processes, certainly the processes that inform the President of the various pieces of information that he requires to make a decision. We know that there is a very narrow window of around 30 minutes for the President to decide in response to something like a Russian ICBM missile, or as little as 15 minutes to respond to a submarine ballistic missile. Again, perhaps AI, given this compressed decision-making timeframe, could help to clarify some of the issues.

The last point is quite reasonable and persuasive. One of the goals of the Act that is stated by the Bill's proposers is that it seeks similar commitments by Russia and China. It appears to be failing quite significantly. According to open sources, both Russia and China seem to be rapidly integrating AI into their nuclear and conventional as well as their dual-use command and control systems. In a way, this eliminates even the possibility that AI could be used in US nuclear systems and, thus, removes any real incentive to bring Russia or China to the negotiation table to discuss how to perhaps reduce these risks.

Q136 **The Lord Bishop of Coventry:** On the back of those fascinating and very helpful comments, I want to turn to Chris King in particular on this question. You were indicating that the NPT in 2026 might consider amendment along these lines to require some form of human involvement. If that is what you were saying, has any movement already been made in that direction by way of preparation? Do you hold out any hope that it might get anywhere, particularly on the back of Dr Johnson's last comments?

Christopher King: First, it would not be an amendment to the treaty as such; it would, rather, be a commitment by the states in question to undertake it. At its review conferences, when they are successful, the NPT usually produces a range of commitments that state parties will undertake over the course of the next review cycle.

The issue has been raised in track 1, 1.5 and 2 risk reduction dialogues on the matter. Certainly, the issue of the nexus between technological developments and nuclear weapons, and the challenges posed therein, was raised during the last review conference. I am hopeful that we can have a dialogue on it during the current review cycle. However, based on what Dr Johnson has said, and I would share some of his concerns, I would not be overly optimistic at this point in time. The review conference is in 2026 and we may be able to make some gains before then.

Dr James Johnson: In terms of the NPT, again this is not my specialty, so I am not really an authority to give any definitive comments here, but certainly I can make a few points about the non-proliferation regime and the non-proliferation treaty itself. There has been some talk that we could have some lessons learned from the NPT structure, and these go through to what we are seeing in the nuclear domain and the intersection of nuclear weapons with artificial intelligence, which is still quite a nascent position.

I would certainly echo Chris's points on the discussions themselves. Perhaps the addition of artificial intelligence to the NPT regime is slightly unrealistic, given the current status.

The other issue to mention from a technical point of view is that AI is very much an evolving technology. Definitions there are still very much up in the air, as we discussed in previous sessions. Nailing down what we mean by "artificial intelligence" as well as the nature of AI, and perhaps what it is capable of doing, how it might be used by adversaries and how it could be calibrated in diverging degrees by different adversaries at different stages all pose challenges for the NPT as a potential model.

In addition, we are talking about a virtual notion of artificial intelligence, which is perhaps more about software than it is about hardware, whereas in the nuclear domain we are talking about a continuous physical phenomenon that involves nuclear fissile materials, for example, which is very much an immutable notion and physical concept that can be verified and monitored, albeit it is quite difficult, given the current challenges.

This poses additional challenges for artificial intelligence, which is continually shifting. It is important to mention that it is fundamentally a dual-use technology. Again, the funding and most of the know-how and technology are coming increasingly from the private sector. They bring with that their own vested interests, and this would bring into notion the increasing amount of stakeholders that would be required to be involved in any non-proliferation discussions.

Q137 **Lord Mitchell:** The Treaty on the Non-Proliferation of Nuclear Weapons does not define nuclear weapons. How has this lack of technical definition affected enforcement? Could lessons be learned regarding a treaty on lethal autonomous weapons?

Christopher King: Just to clarify on the last question, when it comes to non-proliferation I was referring specifically to the Nuclear Non-Proliferation Treaty and the nexus between human control over nuclear weapons or artificial intelligence, and nuclear weapons within the construct of the NPT. The broader non-proliferation/arms control of AI is a bit outside of my purview.

Specifically on your question, the absence of a technical definition of nuclear weapons has not affected enforcement of the treaty. In fact, the absence of a definition is seen to be positive, so much so that the treaty that was concluded two years ago on the prohibition of nuclear weapons also does not include a definition, for the same reasons. Although there have been issues related to compliance, these are largely related to fissile material production. States that have developed nuclear weapons either are or were outside the NPT, or have claimed to have withdrawn from it.

Regarding lessons for LAWS, I note that an internationally agreed definition of LAWS does not exist, and one of the general points of convergence among states at the GGE, under the auspices of the Convention on Certain Conventional Weapons, seems to be that a definition, characterisation or description of LAWS should be technology-neutral, so that it can adequately cover all future developments. Such a definition would describe the functions and capacities of the weapon or weapon systems, and would centre on the element of autonomy of human control—terms that have been at the core of discussions at the GGE. On this point, the GGE already agreed in 2019 that human responsibility for decisions on the use of weapon systems must be retained, since accountability cannot be transferred to machines. This should be considered across the entire lifecycle of the weapon system.

It is important for states to have at least a clear understanding domestically of what they consider to be autonomous weapon systems or lethal autonomous weapon systems, so that they can ensure compliance with international humanitarian law and international human rights law in any potential development, testing or use. However, a rigid definition at the domestic level may later become an obstacle to joining an international instrument on prohibitions and regulations of autonomous weapon systems or lethal autonomous weapon systems and, therefore, are not advisable.

Dr James Johnson: The point that I would like to stress from the outset is that, despite the definitional ambiguities that exist and have caused various loopholes, the NPT itself has been widely ratified and has significantly influenced nuclear policy worldwide. This is hopeful, albeit fiddly, as a signal for future agreements on issues like lethal autonomous weapons.

Although the NPT does not explicitly define nuclear weapons, it clearly delineates the responsibilities and commitments of nuclear weapon states and non-nuclear weapon states, so it does offer a de facto understanding of what constitutes a nuclear weapon. There are some definitions there, albeit quite ambiguous.

On enforcement, there are strengths and weaknesses. We can categorise the main effects into three or four areas. The lack of clear definition can lead to ambiguities and, as I mentioned, loopholes that certain nations can exploit to develop their own nuclear technology that certainly skirt the edges of the treaty. For example, the NPT does not cover things like delivery systems such as missiles, as well as the all-important tactical or non-strategic nuclear weapons, including non-strategic conventional weapons, so things like AI-enhanced cyber weapons, as well as hypersonics and counterspace, which all come into the broader nuclear deterrence rubric. The NPT was drafted in a very different era, when the technology behind nuclear weapons was less advanced than in the current digital age. This lack of definition makes it very challenging to apply the treaty to new technologies and advances in nuclear weaponry.

On verification and compliance, a lack of definition here of nuclear weapons makes it even more challenging to monitor compliance and enforce the treaty. Where one nation, for example, considers that there is a breach, another might argue that they are still within the bounds of the treaty. If we are looking at AI systems, for example, the various cognitive attributes of the systems that exist are not clearly observable. Smart weapons may look like a dumb weapon of the same kind, and it is often just a question of tweaking the software. For example, an autonomous vehicle's sensors that perceive its environment may be visible, but the algorithm is non-visible.

On the second part of your question about the various lessons that we can learn from the NPT, there are several that we can learn for a potential treaty on lethal autonomous weapons. Clearly, tighter definitions are required, which would help to plug any loopholes. A potential LAWS treaty should clearly define what constitutes a lethal weapon, the circumstances for use, when and if it is at all lawful to use them, and whether the weapon is offensive or defensive in nature.

On compliance and verification, a specific international body would need to be tasked with inspections and verifications, so perhaps something like the International Atomic Energy Agency's role in verifying the NPT is a potential mirror or model of the things required to enforce a potential LAWS treaty.

Grey areas would need to be explicitly addressed, such as things like semi-autonomous weapons, mixed mode systems, the degree of human oversight, and specifically where the handoff begins and ends in this very blurred human-machine decision-making continuum. As the technology advances, the treaty needs to be evolved and accommodate future technological developments.

Finally, the balance needs to be quite clear between regulation and progress. Just as the NPT acknowledges the right to use nuclear technology for peaceful purposes, any treaty on lethal autonomous weapons should also recognise its beneficial uses of autonomous technology that do not need to be lethal—non-lethal defence systems for triage assistance for the injured, ISR missions, and verification and testing purposes for nuclear weapons.

In short, creating an international treaty on LAWS presents a considerable challenge, due to the rapid pace of technological change and the diverse range of potential applications that could be involved. That said, behavioural arms control in advance of a law or a treaty could lay the groundwork for such efforts, until such time as geopolitical conditions improved—things like stigmatising bad behaviour such as cyberattacks on NC3 systems. In other words, arms control does not need to be mutually exclusive from strategic arms control, as it has been in the past.

The Chair: Thank you. We are getting to the end of our allotted time. Baroness Hodgson has a concluding question, which is especially designed to be able to be answered in one sentence.

Q138 **Baroness Hodgson of Abinger:** Thank you both so much. If you had one recommendation to make to the UK Government, what would it be?

Christopher King: This is a difficult question to answer, because there are a lot of recommendations that I would like to make to the UK, but coming from the perspective of the United Nations and from a multilateral perspective, the recommendation would be to work with other nuclear armed states, but especially the NPT nuclear weapon states, to develop practical and tangible risk-reduction measures that include the maintenance of human control over the launch of nuclear weapons. A sub-component of this would be to sponsor outreach to all NPT state parties about the significance and consequences of this particular issue.

Dr James Johnson: Thank you for your question. I am glad that it was one recommendation rather than one sentence, as that would be quite tough for an academic, but I will do my best. I have had a good look through the MoD's defence AI strategy as well as the *Ambitious, Safe and Responsible* document, and these are commendable starts. However, as it stands, the strategy report reads very much like an integrated report or review, rather than a strategy that lays out clearly defined choices and priorities. That is to say that it is more of an aspirational rather than operational document.

I would like to see from this document or future documents on strategy more practical guidance on the development of the use of AI NC3 systems, which is one area that the report explicitly said they are considering, as well as how this augmentation can be reconciled also with their AI-free nuclear pledge and what safety measures, as well as ethical and legal standards, are required.

The Government also need to invest in and promote research and consultation on the ethical, security and strategic implications of AI in nuclear weapon systems, as well as, as I mentioned, non-nuclear or conventional counterforce weapons, which are an important part of the conversation. The ultimate goal of this research should be to inform the development of clear and robust policies and guidelines for the use of AI in this context, with a particular focus on preventing accidents, inadvertent escalation and the use of these technologies.

The Chair: Thank you very much indeed for giving us your time this afternoon and for coping, Chris King, in your case, so effectively with technological challenges, which may well have a resonance for the main subject that we are investigating. The one thing that Chris King and I can agree on is that the weather forecast looks very good for Edgbaston for the next five days. Thank you very much.