HOUSE OF LORDS

AI in Weapon Systems Committee

Report of Session 2023–24

Proceed with Caution: Artificial Intelligence in Weapon Systems

Ordered to be printed 23 November 2023 and published 1 December 2023

Published by the Authority of the House of Lords

HL Paper 16

AI in Weapon Systems Committee

The AI in Weapon Systems Committee is appointed to consider the use of artificial intelligence in weapon systems.

Membership

The Members of the Artificial intelligence in weapon systems were:

Lord Browne of Ladyton
Lord Clement-Jones
Baroness Doocey
Lord Lisvane (Chair)
Lord Fairfax of Cameron
Lord Grocott
Lord Hamilton of Epsom
Baroness Hodgson of Abinger
Lord Houghton of Richmond
Lord Mitchell
Lord Sarfraz
Lord Triesman

Declaration of interests

See Appendix 1.

A full list of Members' interests can be found in the Register of Lords' Interests: https://members.parliament.uk/members/lords/interests/register-of-lords-interests

Publications

All publications of the Committee are available at: https://committees.parliament.uk/committee/646/publications/

Parliament Live

Live coverage of debates and public sessions of the Committee's meetings are available at: http://www.parliamentlive.tv

Further information

Further information about the House of Lords and its Committees, including guidance to witnesses, details of current inquiries and forthcoming meetings is available at: http://www.parliament.uk/business/lords

Committee staff

The staff who worked on this inquiry were Alasdair Love (Clerk), Sarah Jennings (Policy Analyst), Cathy Adams (Legal Adviser) and Stephen Reed (Committee Operations Officer).

Contact details

All correspondence should be addressed to the AI in Weapon Systems Committee, Committee Office, House of Lords, London SW1A 0PW. Telephone 020 7219 4878. Email hlaiweapons@parliament.uk

Social Media

You can follow the Committee on X (formerly Twitter): @HLAIWeapons.

CONTENTS

Page

Overview 3

List of abbreviations 6

Chapter 1: Introduction 7
Box 1: Artificial intelligence 7
Box 2: Automation and autonomy in weapon systems 8
The domestic policy landscape 9
Figure 1: Autonomy spectrum framework 11
Box 3: The Ministry of Defence's Five Ethical Principles for AI in Defence 12
The international policy landscape 13
Background to this inquiry 15
Structure of the Report 16

Chapter 2: Automation, Autonomy and AI 17
Defining AWS 17
Definitions by other states 17
Box 4: Definitions of autonomous system, Autonomous Weapon Systems (AWS) and Lethal Autonomous Weapon Systems (LAWS) by other countries 17
Capturing the characteristics of AWS 19
Challenges in creating a single definition 20
The value of a definition 21
Capabilities of AI systems 23
Box 5: Machine Learning 23
Predictability and reliability 24
Data sets 26
Transparency and Explainability 29
Testing, evaluation, verification and validation 30
Meaningful human control and accountability 32
Box 6: Targeting Directives 33

Chapter 3: AWS and the Battlefield of the Future 37
Box 7: Potential battlefield benefits and risks of AWS 37
Changing paradigms of conflict 38
Casualties 38
Sabotage and counter AWS 40
Use of AWS by non-state actors 42
Box 8: Open-source software 43
Arms control 45
Escalation and an AI arms race 46
"Arms race afoot"? 46
Stuck on the escalator? 47
AI in nuclear command and control 48
Box 9: Block Nuclear Launch by Autonomous Artificial Intelligence Act 49
Nuclear escalation 50
Technological risks 51

Chapter 4: International Law 53
AWS and international law 53
Box 10: Basic Principles of international humanitarian law 53
Compliance of AWS with IHL 54
Meaningful human control 56
IHL obligations during weapons development 56
Enforcement and accountability 59
Individual accountability 59
Does the regulation of AWS require new international law? 60
Box 11: Guiding Principles agreed at the 2019 Meeting of the Contracting Parties to the CCW Convention 61

Chapter 5: UK Domestic Policy on AWS 67
Translating strategy to operational practice 67
Procurement, Innovation and Talent 69
Ethics 75
Democratic control 78

Summary of conclusions and recommendations 79

Appendix 1: List of Members and declarations of interest 83

Appendix 2: List of witnesses 86

Appendix 3: Call for evidence 91

Appendix 4: Examples of AWS and AI-enabled weapons 93

Appendix 5: Glossary 96

Evidence is published online at https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/ and available for inspection at the Parliamentary Archives (020 7219 3074).

Q in footnotes refers to a question in oral evidence

OVERVIEW

Artificial intelligence (AI) has spread into many areas of life,[1] and defence is no exception. AI has applications across the military spectrum, from optimising logistics chains to processing large quantities of intelligence data. There is a growing sense that AI will have a major influence on the future of warfare, and forces around the world are investing heavily in capabilities enabled by AI. Despite these advances, fighting is still largely a human activity.

Bringing AI into the realm of warfare through the use of AI-enabled autonomous weapon systems (AWS) could revolutionise defence technology and is one of the most controversial uses of AI today. There has been particular debate about how autonomous weapons can comply with the rules and regulations of armed conflict which exist for humanitarian purposes.

The Government aims to be "ambitious, safe, responsible".[2] Although of course we agree in principle, aspiration has not lived up to reality. In this Report, we therefore make proposals to ensure that the Government approaches development and use of AI in AWS in a way that is ethical and legal, providing key strategic and battlefield benefits, while achieving public understanding and democratic endorsement. "Ambitious, safe and responsible" must be translated into practical implementation.

The Government must seek, establish and retain public confidence and democratic endorsement in the development and use of AI generally, and especially in respect of AWS. It is clear from media coverage of our inquiry that there is widespread interest in and concern about the use of AI in AWS. Achieving democratic endorsement will have several elements:

Understanding: discussion of autonomous weapons and to a significant extent AI in general, is bedevilled by the pursuit of agendas and a lack of understanding. One of our aims is to provide a factual basis for constructive debate, and frankness and transparency on the part of Government will support this process.

The Role of Parliament: Parliament is at the centre of decision-making on the development and use of AWS. Parliament's capacity for oversight depends on the availability of information, on its ability to anticipate issues rather than reacting after the event, and on its ability to hold ministers to account. The Government must allow sufficient space in the Parliamentary timetable and provide enough information for Parliament, including its select committees, to scrutinise its policy on AI effectively. We naturally understand that elements of policy development may be highly sensitive, but there are established ways of dealing with such information. Arguments of secrecy must not be used to sidestep accountability.

Retaining Public Confidence: We are disappointed by the fact that the Ministry of Defence does not "currently undertake monitoring or polling to understand

[1] We note the risk of over trusting technology, as demonstrated by the Post Office Horizon IT Scandal.

[2] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]

public attitudes towards the use of autonomous weapons systems".[3] The Government must ensure that it properly consults the public on the development of AWS. It must also ensure ethics are at the centre of its policy, including expanding the role of the Ministry of Defence's AI Ethics Advisory Committee.

Crucial to this process will be achieving the following aims:

The Government should lead by example in international engagement on regulation of AWS. The AI Safety Summit was a welcome initiative, but it did not cover defence. The Government must include AI in AWS in its proclaimed desire to "work together in an inclusive manner to ensure human-centric, trustworthy and responsible AI that is safe" and to support "the good of all through existing international fora and other relevant initiatives".[4]

The international community has been debating the regulation of AWS for several years. Outcomes from this debate could be a legally binding treaty or non-binding measures clarifying the application of international humanitarian law—each approach has its advocates. Despite differences about form, the key goal is accelerating efforts to achieving an effective international instrument.

A key element in this will be prohibiting the use of AI in nuclear command, control and communications. On one hand, advances in AI have the potential to bring greater effectiveness to nuclear command, control and communications. For example, machine learning could improve detection capabilities of early warning systems, make it easier for human analysts to cross-analyse intelligence, surveillance and reconnaissance data, and improve the protection of nuclear command, control and communications against cyberattacks.

However, use of AI in nuclear command, control and communications also has the potential to spur arms races or increase the likelihood of states escalating to nuclear use—either intentionally or accidentally—during a crisis. The compressed time for decision-making when using AI may lead to increased tensions, miscommunication, and misunderstanding. Moreover, an AI tool could be hacked, its training data poisoned, or its outputs interpreted as fact when they are statistical correlations, all leading to potentially catastrophic outcomes.

The Government should adopt an operational definition of AWS. Surprisingly, the Government does not currently have one. The Ministry of Defence has stated it is cautious about adopting one because "such terms have acquired a meaning beyond their literal interpretation" and is concerned that an "overly narrow definition could become quickly outdated in such a complex and fast-moving area and could inadvertently hinder progress in international discussions".[5] However, we believe it is possible to create a future-proofed definition. Doing so would aid the UK's ability to make meaningful policy on autonomous weapons and engage fully in discussions in international fora.

[3] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[4] DSIT, Foreign Commonwealth and Development Office, and Prime Minister's Office, 'The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023' (1 November 2023): https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [accessed 22 November 2023]

[5] Written evidence from the Ministry of Defence (AIW0035)

The Government should ensure human control at all stages of an AWS's lifecycle. Much of the concern about AWS is focused on systems in which the autonomy is enabled by AI technologies, with an AI system undertaking analysis on information obtained from sensors. But it is essential to have human control over the deployment of the system both to ensure human moral agency and legal compliance. This must be buttressed by our absolute national commitment to the requirements of international humanitarian law.

The Government should ensure that its procurement processes are appropriately designed for the world of AI. We heard that the Ministry of Defence's procurement suffers from a lack of accountability and is overly bureaucratic. In particular, we heard that it lacks capability in relation to software and data, both of which are central to the development of AI. This may require revolutionary change. If so, so be it; but time is short.

In this Overview we have set out the principal themes of this Report. They are underpinned by our detailed recommendations in the chapters that follow.

LIST OF ABBREVIATIONS

AI	Artificial Intelligence
AWS	Autonomous Weapon System(s)
CCW	UN Convention on Certain Conventional Weapons
CIWS	Close-in Weapon System
DCMS	Department for Digital, Culture, Media and Sport
DSIT	Department for Science, Innovation and Technology
FCDO	Foreign, Commonwealth and Development Office
GGE	Group of Governmental Experts (established by states parties to the CCW)
ICRC	International Committee of the Red Cross
IHL	International Humanitarian Law
IR	Integrated Review of Security, Defence, Development and Foreign Policy
ISR	Intelligence, surveillance and reconnaissance
LAWS	Lethal Autonomous Weapon System(s)
LLM	Large Language Model
ML	Machine Learning
MoD	Ministry of Defence
NATO	North Atlantic Treaty Organisation
NPT	Nuclear Non-Proliferation Treaty
RL	Reinforcement Learning
SMEs	Small and Medium-sized Enterprises
TTP	Tactics, Techniques and Procedures
UN	United Nations

Proceed with Caution: Artificial Intelligence in Weapon Systems

CHAPTER 1: INTRODUCTION

1. The use of artificial intelligence (see Box 1) for defence and security purposes is among the most emotive and high-stakes areas of AI development today.[6] Some aspects are less controversial, such as the use of AI for non-violent defence applications, including general data analysis, cyber defence, intelligence gathering, surveillance and reconnaissance (ISR), predictive maintenance and logistics. Many of these applications, in particular cyber defence, take place in the 'grey zone': competition among state and non-state actors that falls between traditional war and peace.[7]

2. More contentiously, and the focus of this Report, AI is used to enable certain autonomous weapon systems (AWS) that can select, detect and engage targets with little to no human intervention or possess some degree of autonomy in one or more aspects (see Box 2).[8] Neither autonomy nor the use of AI in weapon systems is especially new,[9] although the proliferation of AI within AWS is.

Box 1: Artificial intelligence

AI lacks a universally agreed definition, but the Ministry of Defence defines it "as a family of general-purpose technologies, any of which may enable machines to perform tasks normally requiring human or biological intelligence, especially when the machines learn from data how to do those tasks."

AI can typically be viewed as being either 'general' or 'narrow' in scope. Artificial general intelligence refers to a machine with broad cognitive abilities, which is able to perform, or at least simulate convincingly, many of the intellectual capacities of a human being, and potentially surpass them. By contrast, narrow AI systems perform specific tasks which would require intelligence and may even surpass human abilities in these areas. However, such narrow systems are limited in the range of tasks they can perform.

Although recent generative AI models have begun to display general capabilities, there are different views about how long it will take to develop artificial general intelligence. This Report therefore deals almost exclusively with narrow AI. Similarly, technologies which have the potential to change AI in the future but which are further off—such as quantum computing—are not discussed.

Source: MoD, Defence Artificial Intelligence Strategy (June 2022): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 10 August 2023] and Artificial Intelligence Committee, AI in the UK: ready, willing and able? (Report of Session 2017–19, HL Paper 100).

[6] This was suggested by our predecessor, the Artificial Intelligence Committee, *AI in the UK: ready, willing and able?* (Report of Session 2017–19, HL Paper 100), para 334

[7] MoD, 'Getting to grips with grey zone conflict' (26 April 2021): https://stratcommand.blog.gov.uk/2021/04/26/getting-to-grips-with-grey-zone-conflict/ [accessed 2 August 2023]

[8] It should be noted that autonomous weapons, such as anti-personnel mines, can exist independent of what is typically described as AI.

[9] Q 97 (Professor Durrant-Whyte)

Box 2: Automation and autonomy in weapon systems

Automation refers to the use of systems to perform tasks that would ordinarily involve human input. Automation and autonomy can be viewed as existing on a spectrum relating to the level of human supervision over a system. This can range from manually controlled systems to those that independently make decisions about how to achieve certain human-set goals. AI technologies are the primary enabler of autonomy.

The US National Institute of Standards and Technology defines a weapons system as "a combination of one or more weapons with all related equipment, materials, services, personnel, and means of delivery and deployment (if applicable) required for self-sufficiency."

Autonomous weapon systems (AWS) are weapon systems which can select, detect and engage targets with little to no human intervention, or which possess some degree of autonomy in one or more respects. The scope of these systems can vary significantly, from fully autonomous weapons that can operate without any human involvement, to partially autonomous weapons that require human action to launch an attack.

For clarity, 'AWS' is used throughout this Report. This is to ensure that systems which are not designed to engage targets are included, although those which are capable of using 'lethal' force are our focus. Certain unarmed systems (for example, software-based decision support tools) may play a key role in identifying targets.

Source: NIST Computer Security Resource Centre, 'Weapons System': https://csrc.nist.gov/glossary/term/weapons_system [accessed 24 November 2023]. Betrand Meyer, 'John McCarthy', Communications of the ACM (28 October 2011): https://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext [accessed 24 November 2023].

3. **Discussion of AWS and, to a significant extent, AI is bedevilled by lack of understanding, misunderstanding, and the pursuit of agendas. A Select Committee is ideally placed to develop independent and impartial analysis in order to provide a sound basis for well-informed public debate. That has been our principal purpose in this inquiry. Specifically, in this Report we aim to answer seven key questions:**

- **What is an AWS, and should the Government define it?**
- **What are the technological abilities and limitations of the AI enabling AWS?**
- **What ethical principles should apply to use of AWS and how can they be implemented?**
- **To what extent is human involvement in the operation of AWS desirable and necessary for compliance with international humanitarian law?**
- **What are the benefits and risks associated with using AI in AWS?**
- **How should the UK engage internationally on regulation of AWS?**
- **Are the Ministry of Defence's internal and national policies fit for purpose?**

4. AI has been an area of rapid development in recent years, with narrow AI finding applications in many areas of life. One of the founders of the field of AI, John McCarthy, once said "As soon as it works, no one calls it AI any more."[10] This difficulty in creating a static definition of AI makes it challenging to set clear boundaries on what forms and applications of AI are acceptable.

5. This fast pace of development, as well as the lack of publicly available information on how AI is being developed, also pose issues for Parliamentary scrutiny. The Right Hon Lord Sales, Justice of the Supreme Court, has offered a stark assessment of the impact of AI on democracy:

> "Through lack of understanding and access to relevant information, the power of the public to criticise and control the systems which are put in place to undertake vital activities in both the private and the public sphere is eroded. Democratic control of law and the public sphere is being lost."[11]

6. It is clear from the extensive press coverage of our inquiry that there is wide public interest in and concern about the use of AI in weapon systems. It is therefore essential that Parliament is at the centre of decision-making on their development and use. Parliament's capacity for oversight depends on transparency and availability of information, on its ability to anticipate issues rather than reacting after the event, and on its ability to hold ministers to account. We aim to contribute to this process through this Report.

7. The capabilities and potential of AI may well affect almost every area of human activity, including the exercise of military power and the conduct of armed conflict. Professor Sir Michael Howard characterised the changed world after 1945 as one in which "war is now seen as being a matter for governments and not for peoples; an affair of mutual destruction inflicted at remote distances by technological specialists operating according to the arcane calculations of strategic analysts".[12] The development and application of AI could produce a 21st-century version of such seismic change.

The domestic policy landscape

8. The Government asserts that AI specifically, and science and technology more broadly, is at the heart of its defence strategy. We take this to mean primarily their development of defence capability. In March 2023, the Government published its Integrated Review of Security, Defence, Development and Foreign Policy Refresh, updating the 2021 review (IR2021). In the 2023 Review the Government says that "We will build on IR2021's prioritisation of strategic advantage in science and technology as a core national priority".[13]

[10] Communications of the ACM, 'John McCarthy' (28 October 2011): https://cacm.acm.org/blogs/blog-cacm/138907-john-mccarthy/fulltext [accessed 22 November 2023]

[11] Lord Sales, 'Algorithms, Artificial Intelligence and the Law', Judicial Review, vol. 25 (June 2020), p 51: https://www.tandfonline.com/doi/full/10.1080/10854681.2020.1732737?needAccess=true [accessed 22 November 2023]

[12] Michael Howard, "Empires, Nations and Wars": the Yigal Allon Memorial Lecture, 1982

[13] HM Government, *Global Britain in a competitive age: Integrated Review of Security, Defence, Development and Foreign Policy*, CP 403 (March 2021), p 35: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/975077/Global_Britain_in_a_Competitive_Age-_the_Integrated_Review_of_Security__Defence__Development_and_Foreign_Policy.pdf [accessed 22 November 2023]

9. The 2023 Review stated that "We are a top five nation in innovation, artificial intelligence (AI) and cyber, and a major international power in science and technology. We will increase our resilience for the long term by surging investment into these areas."[14] In 2021, the UK spent $2.7 billion on defence research and development (some of which will be on AI in defence). This compares to $76.4 billion by the US, $2.3 billion by Germany, $2.6 billion by France, and $1.1 billion by Japan.[15] UK statistics for spending on AI in defence specifically are not available.

10. Alongside the publication of the 2021 Integrated Review, the Ministry of Defence published *Defence in a competitive age*, setting out the current and future challenges faced in defence and the intentions to tackle them. It stated that AI will be "essential" to modernising defence across the board and that "future conflicts may be won or lost on the speed and efficacy of the AI solutions employed."[16]

11. In June 2022, the Ministry of Defence published its Defence AI Strategy.[17] The Strategy establishes the Ministry of Defence's intention to utilise AI from "'back office' to battlespace" and clarifies that this includes AI-enabled weapon systems. The Strategy established four key objectives:

- Transform the Ministry of Defence into an 'AI-ready' organisation;
- Adopt and exploit AI at pace and scale[18] for defence advantage;
- Strengthen the UK's defence and security AI ecosystem; and
- Shape global AI developments to promote security, stability and democratic values.[19]

12. The Defence AI Strategy sets out an "autonomy spectrum framework" which establishes differing levels of human input in autonomous systems (Figure 1). However, to some extent this figure simplifies human control. There are different levels of interrelationship between humans and machines, rather than a simple ascending scale of human involvement. The level of human involvement, whether in oversight, verification or control, will vary depending upon the design of the system, the mission objectives and the operational context of where and how the AI system is being used.

[14] HM Government, *Integrated Review Refresh 2023: Responding to a more contested and volatile world*, CP 811 (March 2023), p 4: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1145586/11857435_NS_IR_Refresh_2023_Supply_AllPages_Revision_7_WEB_PDF.pdf [accessed 22 November 2023]

[15] OECD, 'Main Science and Technology Indicators': https://stats.oecd.org/Index.aspx?DataSetCode=MSTI_PUB [accessed 20 October 2023]

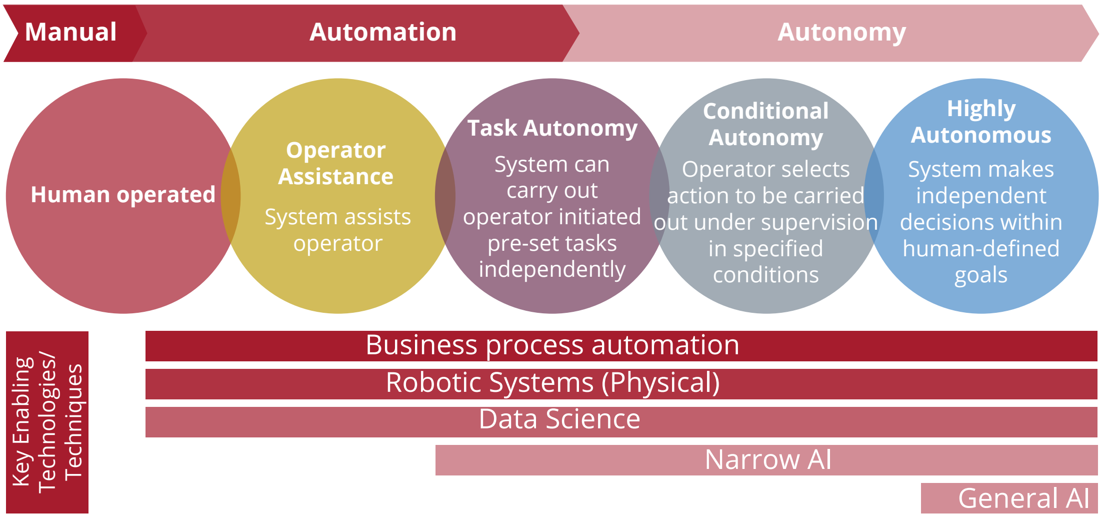
[16] MoD, *Defence in a competitive age*, CP 411 (March 2021), p 42: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/974661/CP411_-Defence_Command_Plan.pdf [accessed 10 August 2023]

[17] MoD, *Defence Artificial Intelligence Strategy* (June 2022): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 10 August 2023]

[18] Which we understand to mean "quickly and extensively".

[19] MoD, *Defence Artificial Intelligence Strategy*, pp 6–7

Figure 1: Autonomy spectrum framework


Manual
Automation
Autonomy
Human operated
Operator Assistance
System assists operator
Task Autonomy
System can carry out operator initiated pre-set tasks independently
Conditional Autonomy
Operator selects action to be carried out under supervision in specified conditions
Highly Autonomous
System makes independent decisions within human-defined goals
Key Enabling Technologies/ Techniques
Business process automation
Robotic Systems (Physical)
Data Science
Narrow AI
General AI

Source: MoD, Defence Artificial Intelligence Strategy (June 2022), p 4: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 10 August 2023]

13. The Defence AI Strategy gives limited consideration to the specific issue of AWS. The Strategy states that "We are clear that we seek to maximise our operational capability through the use of AI, but also that there must be no 'race to the bottom'–no pursuit of capability without regard for responsibilities and safeguards."[20] It asserts the Government's view that international humanitarian law (IHL) is sufficient to regulate AWS and that "nothing about AI fundamentally changes our obligations under UK law and international law, or the importance we attach to the standards, values and norms of the society we serve."[21] It states that the Government is open to mechanisms to promote best practice, including codes of conduct, positive obligations or commitments, or reporting or verification mechanisms.[22]

14. Alongside the Defence AI Strategy, the Ministry of Defence published a separate policy statement on the ethical and responsible application of AI in defence, called *Ambitious, safe, responsible*.[23] The statement recognises key challenges in adopting AI for defence, including bias in datasets[24] and the impact of AI on responsibility and accountability (both discussed in Chapter 2).[25] The statement also sets out an ethical framework to guide the application of AI to defence alongside an AI Ethics Advisory Panel to scrutinise the Ministry of Defence's work on establishing responsible and ethical AI (see Box 3).[26] The Defence AI Strategy and ethics principles are discussed further in Chapter 5.

[20] MoD, Defence Artificial Intelligence Strategy, p 52
[21] *Ibid.*
[22] *Ibid.*, p 54
[23] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]
[24] Bias occurs when certain types of data are missing or more represented than others, often deriving from how the data was obtained or sampled. Biases may result in unrepresentative or undesirable outputs.
[25] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence*
[26] *Ibid.*

Box 3: The Ministry of Defence's Five Ethical Principles for AI in Defence

- **Human Centricity:** consideration of the impact of any AI systems on humans throughout the lifecycle of the system.
- **Responsibility**: establishing human responsibility and accountability for AI-enabled systems.
- **Understanding**: ensuring that relevant individuals appropriately understand AI-enabled systems and their outputs.
- **Bias and harm mitigation**: requiring those responsible for AI-enabled systems to proactively mitigate risk and biases from the systems.
- **Reliability**: AI-enabled systems must be demonstrably reliable and secure.

Source: MoD, Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]

15. However, just as the Defence AI Strategy does not focus on the application AWS, *Ambitious, safe, responsible* is limited to one page and an annex on AWS.[27] As stated in the Defence AI Strategy, the Ministry of Defence has not ruled out using AI in weapon systems but specifies that AI weapon systems which identify, select and attack targets must have "context-appropriate human involvement".[28] This term is discussed further in Chapters 2 and 4. The statement reiterates that the UK does not have fully autonomous weapon systems and does not intend to develop them.[29] The statement sets out that the Department believes that "AI within weapon systems can and must be used lawfully and ethically."[30]

16. On 18 July 2022 the Department for Digital, Culture, Media and Sport published a policy paper on *Establishing a pro-innovation approach to regulating AI.*[31] The paper establishes a context-specific approach, which it anticipates will enable sectors like defence, which have distinct approaches to AI, to continue to develop the regulatory mechanisms needed based on the context.[32] In March 2023 the Department for Science, Innovation and Technology and the Foreign, Commonwealth and Development Office published an International Technology Strategy, setting out the principles governing the UK's international engagement on technology and how it is expected to shape global AI governance.[33] In August 2023 the Department for Science, Innovation and Technology published the AI governance white paper, which sets out the Government's plans to take a "pro-innovation approach" to become a "science and technology superpower by 2030", while also leading

[27] *Ibid.*

[28] *Ibid.*

[29] However, by our definition, Phalanx is a fully autonomous weapon system (see para 55).

[30] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence*

[31] DCMS, *Establishing a pro-innovation approach to regulating AI An overview of the UK's emerging approach*, CP 728 (July 2022): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1092630/_CP_728__-_Establishing_a_pro-innovation_approach_to_regulating_AI.pdf [accessed 22 November 2023]

[32] *Ibid.*, p 11

[33] DSIT and FCDO, 'The UK's International Technology Strategy' (March 2023): https://www.gov.uk/government/publications/uk-international-technology-strategy/the-uks-international-technology-strategy [accessed 10 August 2023]

"the international conversation on AI governance and demonstrate[ing] the value of our pragmatic, proportionate regulatory approach."[34]

17. **The lack of available statistics on the UK's spending on AI in defence means that it is difficult to determine whether the level of spending is appropriate and to compare it internationally.** ***The Government must publish annual spending on AI in defence as part of the Ministry of Defence's Finance and Economics Statistics Bulletin series.***

The international policy landscape

18. There are no international legal prohibitions on AWS *per se*. However, IHL sets limits on the development and use of weapons on the battlefield.[35] These rules apply to all weapons, AWS included. The Government's position is that IHL "provides a robust, principle-based framework for the regulation of weapons development and use" and it remains "the most appropriate way of regulating new means and methods of warfare."[36]

19. Attempts to regulate AWS have primarily been undertaken through the United Nations (UN). The states parties to the Convention on Certain Conventional Weapons (CCW) established a Group of Governmental Experts in 2017 to discuss emerging technologies in the area of lethal autonomous weapon systems (LAWS). The Group of Governmental Experts most recently met in May 2023, where the Government reaffirmed its position that international humanitarian law is sufficient to regulate AWS. The UK joined Australia, Canada, Japan, the Republic of Korea and the United States in setting out a position that AWS must not be designed to:

- Target civilians or civilian objects, or to spread terror among the civilian population;
- Conduct engagements that would invariably result in incidental loss of civilian life, injury to civilians, and damage to civilian objects excessive in relation to the concrete and direct military advantage anticipated; or
- Conduct engagements that would not be the responsibility of the commanders and operators using the system.[37]

20. International humanitarian law and the Group of Governmental Experts are discussed further in Chapter 4.

21. The UK Government also engages in international discussion of AI ethics. James Cartlidge MP, Minister for Defence Procurement, told us that technical and capability development collaboration is focused on the USA, Five Eyes

[34] DSIT, 'A pro-innovation approach to AI regulation' (August 2023): https://www.gov.uk/government/publications/ai-regulation-a-pro-innovation-approach/white-paper [accessed 10 August 2023]

[35] International Committee of the Red Cross, *What is International Humanitarian Law* (July 2004): https://www.icrc.org/en/doc/assets/files/other/what_is_ihl.pdf [accessed 22 November 2023]

[36] MoD, *Defence Artificial Intelligence Strategy*, p 52

[37] Convention on Prohibition or Restriction on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Draft articles on autonomous weapon systems - prohibitions and other regulatory measures on the basis of international humanitarian law* (13 March 2023), p 2: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.4_Rev1.pdf [accessed 22 November 2023]

nations[38] and NATO.[39] The Ministry of Defence has also "reached out" to partners such as Germany, France, Japan, India, Korea and Singapore to "explain our approach to responsible AI, build communities of interest and champion core values".[40] In November 2023, the UK Government joined states in endorsing the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy.[41] Among other things, this declaration endorsed the following measures:

- States should take appropriate steps, such as legal reviews, to ensure that their military AI capabilities will be used consistent with their respective obligations under international law.
- States should take proactive steps to minimise unintended bias in military AI capabilities.
- States should ensure that military AI capabilities are developed with methodologies, data sources, design procedures, and documentation that are transparent to and auditable by their relevant defence personnel.
- States should ensure that personnel who use or approve the use of military AI capabilities are trained so they sufficiently understand the capabilities and limitations of those systems in order to make appropriate context-informed judgments on the use of those systems and to mitigate the risk of automation bias.
- States should ensure that the safety, security, and effectiveness of military AI capabilities are subject to appropriate and rigorous testing and assurance within their well-defined uses and across their entire life-cycles.
- States should implement appropriate safeguards to mitigate risks of failures in military AI capabilities, such as the ability to detect and avoid unintended consequences and the ability to respond.[42]

22. On 1–2 November 2023, the Government hosted the AI Safety Summit ("the Bletchley Summit"). This was attended by countries, academics, civil society representatives and companies. The aim of the Summit was to "focus on how to best manage the risks from the most recent advances in AI" and to discuss the need for "an urgent international conversation given the rapid

[38] An intelligence alliance comprising the UK, USA, Australia, Canada and New Zealand.

[39] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence

[40] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/. Notably, China and the USA have recently held discussions on AI safety. See Breaking Defense, 'Biden launces AI 'risk and safety' talks with China. Is nuclear C2 a likely focus?' (15 November 2023): https://breakingdefense.com/2023/11/biden-launches-ai-risk-and-safety-talks-with-china-is-nuclear-c2-a-likely-focus/ [accessed 16 November 2023]

[41] US Department of State, 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (1 November 2023): https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy/ [accessed 16 November 2023]

[42] Liber Institute West Point, 'The Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (13 November 2023): https://lieber.westpoint.edu/political-declaration-responsible-military-use-artificial-intelligence-autonomy/ [accessed 16 November 2023] and US Department of State, 'Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy' (9 November 2023): https://www.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy-2/ [accessed 16 November 2023]

pace at which the technology is developing”.[43] The Government reiterated its commitment to international discussion of AI in the King’s Speech at the beginning of the present session of Parliament.[44]

23. However, the Summit did not cover use of AI in defence. The result of the Summit was the Bletchley Declaration signed by all countries in attendance, including, among others, China, the EU, Israel, and the USA.[45] As well as committing to support “an internationally inclusive network of scientific research on frontier AI safety”, the Declaration stated:

- AI should be designed, developed, deployed, and used, in a manner that is safe, in such a way as to be human-centric, trustworthy and responsible.
- We welcome relevant international efforts to examine and address the potential impact of AI systems in existing fora and other relevant initiatives, and the recognition that the protection of human rights, transparency and explainability, fairness, accountability, regulation, safety, appropriate human oversight, ethics, bias mitigation, privacy and data protection need to be addressed.
- Substantial risks may arise from potential intentional misuse or unintended issues of control relating to alignment with human intent.
- We resolve to work together in an inclusive manner to ensure human-centric, trustworthy and responsible AI that is safe, and supports the good of all through existing international fora and other relevant initiatives, to promote cooperation to address the broad range of risks posed by AI.
- Whilst safety must be considered across the AI lifecycle, actors developing frontier AI capabilities, in particular those AI systems which are unusually powerful and potentially harmful, have a particularly strong responsibility for ensuring the safety of these AI systems, including through systems for safety testing, through evaluations, and by other appropriate measures.

24. ***The Bletchley Declaration of November 2023 is, inevitably, aspirational, but it is a start. We commend the contents of the Declaration and encourage the Government to apply its principles to AI in defence.***

Background to this inquiry

25. The establishment of this Committee was recommended by the Liaison Committee in November 2022. The proposal for a “special inquiry committee to examine the use of artificial intelligence in weapon systems”

[43] DSIT, ‘AI Safety Summit: introduction’ (31 October 2023): https://www.gov.uk/government/publications/ai-safety-summit-introduction/ai-safety-summit-introduction-html [accessed 7 November 2023]

[44] Prime Minister’s Office, ‘The King’s Speech 2023’ (7 November 2023): https://www.gov.uk/government/speeches/the-kings-speech-2023 [accessed 7 November 2023]

[45] DSIT, Foreign Commonwealth and Development Office, and Prime Minister’s Office, ‘The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023’: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 [accessed 7 November 2023]

was made by Lord Clement-Jones, now a Member of this Committee. This followed the House's Select Committee on Artificial Intelligence, chaired by Lord Clement-Jones, which recommended "that the UK's definition of autonomous weapons should be realigned to be the same, or similar, as that used by the rest of the world."[46]

26. During our inquiry, we received 43 written submissions and heard from 35 witnesses in oral evidence sessions. We thank everyone who submitted written evidence and gave oral evidence.

27. We undertook three visits. On 21 June 2023, we visited Cambridge, including RAND Europe Cambridge, the Leverhulme Centre for the Future of Intelligence (hosted by Emmanuel College), and the Information Engineering Division of the University of Cambridge Engineering Department. On 12 to 13 September 2023, we visited Glasgow and Edinburgh, including the University of Strathclyde, the Autonomous Systems and Connectivity Research Division, School of Engineering, University of Glasgow, and the School of Informatics, University of Edinburgh. On 19 September we visited the Permanent Joint Headquarters at Northwood to be briefed on targeting and IHL. We thank everyone who participated in these visits, in particular the students and postdoctoral researchers at the University of Strathclyde for their fascinating presentations on the redesigning of a weapon with an AI-based operating system.

28. We are grateful to our specialist advisers, Dr Adrian Weller, Director of Research in Machine Learning, University of Cambridge, and Professor Dame Muffy Calder, Vice-Principal and Head of College of Science and Engineering and Professor of Formal Methods, University of Glasgow.

Structure of the Report

29. In Chapter 2 we examine what constitutes an AWS and the implications of autonomy derived from AI. We cover evidence we received on the AI models which provide the autonomy underpinning AWS, including an assessment of their abilities and limitations.

30. In Chapter 3 we look at the possible impact of AWS on the battlefield, including the impact on the number and nature of casualties, its use by non-state actors,[47] its impact on the speed of escalation, and AI's role in nuclear command, control and communications.

31. In Chapter 4 we further discuss how IHL applies to AWS, with a summary of efforts to regulate AWS through international fora.

32. In Chapter 5 we examine how the Government has approached development and use of AWS on a domestic level, including the Government's broader position on AI ethics, development and regulation.

33. This Report is not just about the roles and characteristics of AI-enabled AWS—but also those of humans, such as empathy, judgement, morality, responsibility and conscience, which clearly differentiate us from AI.

[46] Artificial Intelligence Committee, *AI in the UK: ready, willing and able?* (Report of Session 2017–19, HL Paper 100), p 101

[47] Organisations or individuals that are not affiliated with governments.

CHAPTER 2: AUTOMATION, AUTONOMY AND AI

34. In considering policy on AWS, it is important to address some key questions:

- What is an AWS and how can it be defined?

- What are the specific challenges of AI-enabled AWS, above and beyond those of other weapon systems?

- How does the autonomous nature of the system change the way human operators interact with and exert control over it, and what implications does this have for questions of accountability?

Defining AWS

35. This section addresses the question of what constitutes an Autonomous Weapon System (AWS), the differing definitions used by states and international organisations, and the challenges and benefits of producing a single workable definition of AWS.

Definitions by other states

36. There is no single accepted definition of what constitutes an AWS; and clarifying the term has been a focus of international policymaking for many years. Many states and international organisations have adopted working definitions of what constitutes an AWS or an autonomous system. Several of these are set out in Box 4.

Box 4: Definitions of autonomous system, Autonomous Weapon Systems (AWS) and Lethal Autonomous Weapon Systems (LAWS) by other countries

AWS: "A weapon system that, once activated, can select and engage targets without further intervention by a human operator. This includes, but is not limited to, operator-supervised autonomous weapon systems that are designed to allow operators to override operation of the weapon system, but can select and engage targets without further operator input after activation." The United States (2023)

"LAWS should include but not be limited to the following 5 basic characteristics. The first is lethality, which means sufficient pay load (charge) and/or means to be lethal. The second is autonomy, which means absence of human intervention and control during the entire process of executing a task. Thirdly, impossibility for termination, meaning that once started there is no way to terminate the device. Fourthly, indiscriminate effect, meaning that the device will execute the task of killing and maiming regardless of conditions, scenarios and targets. Fifthly evolution, meaning that through interaction with the environment the device can learn autonomously, expand its functions and capabilities in a way exceeding human expectations." China (2018)

"The ICRC understands AWS to be weapons that select and apply force to targets without human intervention. After initial activation or launch by a person, an AWS self-initiates or triggers a strike in response to information from the environment received through sensors and on the basis of a generalized "target profile" (technical indicators function as a generalized proxy for a target)." International Committee of the Red Cross (2022)

"Autonomous: pertaining to a system that decides and acts to accomplish desired goals, within defined parameters, based on acquired knowledge and an evolving situational awareness, following an optimal but potentially unpredictable course of action." NATO (2020)

Source: Department of Defence 'Directive 3000.09, Autonomy in Weapon Systems' (25 January 2023), p 21: https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf [accessed 20 September 2023]. Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, 'Non-exhaustive compilation of definitions and characterizations' (CCW/GGE.1/2023/CRP.1, 10 March 2023), p 3: https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_CRP.1_0.pdf [accessed 20 September 2023]. International Review of the Red Cross, 'International Committee of the Red Cross (ICRC) position on autonomous weapon systems: ICRC position and background paper' (IRRC No. 915, January 2022): https://international-review.icrc.org/articles/icrc-position-on-autonomous-weapon-systems-icrc-position-and-background-paper-915 [accessed 20 September 2023]. NATO, AAP-06 Edition 2020: NATO glossary of terms and definitions (2020), p 16: https://www.jcs.mil/Portals/36/Documents/Doctrine/Other_Pubs/aap6.pdf [accessed 20 September 2023]

37. Surprisingly, the UK does not have an operational definition of AWS. The Ministry of Defence has stated it is cautious about adopting one because "such terms have acquired a meaning beyond their literal interpretation" and concerns that an "overly narrow definition could become quickly outdated in such a complex and fast-moving area and could inadvertently hinder progress in international discussions".[48] Instead, it uses the latest definitions of 'autonomous' systems set out by the North Atlantic Treaty Organisation (NATO) (see Box 4), although this leaves ambiguous terms such as "desired, "goals", "parameters" and "unpredictable". However, the UK has previously published non-operative definitions of AWS and signed up to joint submissions with other countries which define AWS. At the 2016 Informal Meeting of Experts on Lethal Autonomous Weapons Systems, the UK set out its understanding of the term AWS as:

> "One which is capable of understanding, interpreting and applying higher level intent and direction based on a precise understanding and appreciation of what a commander intends to do and perhaps more importantly why".[49]

38. Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley, criticised this understanding, noting that it would exclude any weapon that follows an "identifiable set of instructions". He argued that the UK is "using the words "autonomous weapon" in a way that is a much higher standard than the rest of the world, so it is not denying itself anything except for weapons that may never exist".[50]

[48] Written evidence from MoD (AIW0035)

[49] Foreign and Commonwealth Office, *United Kingdom of Great Britain and Northern Ireland Statement to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems* (April 2016): https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Informal_Meeting_of_Experts_(2016)/2016_LAWS%2BMX_Towardaworkingdefinition_Statements_United%2BKindgom.pdf [accessed 22 November 2023]

[50] Q 120

39. More recently, a working paper submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems (GGE) by a group of nations including the UK[51] defined AWS as:

> "Including … novel and more sophisticated weapons with autonomous functions, including those weapon systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator."[52]

40. Sir Chris Deverell, former Commander of the UK's Joint Forces Command, was critical of the UK's failure to adopt a working definition of AWS, saying "there is something slightly odd about having a policy not to have something without defining what it is".[53] Other evidence we have received has criticised the previous definitions presented by the UK Government as setting "futuristic and unrealistic thresholds" which are "almost meaningless".[54] Much of the evidence we received called on the UK to adopt a working definition of AWS.[55]

Capturing the characteristics of AWS

41. Many witnesses felt that it was important for any definition to capture several elements of the system, including:

- Autonomy: The ability of the system to complete the entire engagement cycle with little or no human involvement, from target identification, through to selection, and finally to engagement. [56]

- AI-enabled: The use of AI technologies as the primary enabler of that autonomy, undertaking machine analysis on information obtained from sensors to enable performance of the critical functions of acquiring, tracking, selecting, and attacking military objectives.[57] This enables a system to identify targets that are in line with criteria predefined during the programming phase, but may not be explicitly pre-specified, may be hard for humans to predict, and may include adaptive capabilities, where a system can further 'learn' and adapt its behaviour after deployment.[58]

- Purpose of use: The system is pre-programmed to target types or categories of targets, rather than a specific designated target.[59] The

[51] Other states included Australia, Canada, Japan, Poland, the Republic of Korea and the United States.

[52] Convention on Prohibition or Restriction on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Draft articles on autonomous weapon systems - prohibitions and other regulatory measures on the basis of international humanitarian law*, p 1

[53] Q 144 (Sir Chris Deverell)

[54] Written evidence from Dr Ingvild Bode, Dr Hendrik Huelss and Anna Nadibaidze (AIW0015)

[55] Q 144 (Sir Chris Deverell), written evidence from James Baker, Executive Director (Policy and Operations) and Labour for the Long Term (AIW0031)

[56] Written evidence from Dr Mikolaj Firlej (AIW0034)

[57] Written evidence from Dr Mikolaj Firlej (AIW0034), written evidence from Dr Ingvild Bode, Dr Hendrik Huelss and Anna Nadibaidze (AIW0015), written evidence from Dr Ozlem Ulgen (AIW0019)

[58] Mariarosaria Taddeo and Alexander Blanchard, Science and Engineering Ethics, *A Comparative Analysis of the Definitions of Autonomous Weapons Systems* (23 August 2022): https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9399191/ [accessed 22 November 2023]

[59] Q 19 (Dr Vincent Boulanin)

system is deployed with the purpose of applying destructive (whether anti-material or lethal) force.[60]

42. Tsvetelina van Benthem of the University of Oxford noted that autonomy exists on a spectrum, with varying types and degrees of autonomy possible, and that autonomy can exist in relation to a range of functions such as the assessment of data to identify targets or autonomy in decisions to apply force.[61] She noted that a definition of AWS may look less like the singular definition provided for a chemical weapon or cluster munition, but may instead entail "some form of common understanding on characteristics and elements."[62] Similarly, Dr David Anderson, Reader in Autonomous Systems and Connectivity, University of Glasgow, said that a definition should "make it crystal clear" that "autonomy is not binary — autonomy lies on a spectrum which goes from human operation (minimal autonomy) through automatic systems (partial autonomy) on to fully-autonomous systems".[63]

43. While discussions of such systems have often focused on the issue of autonomy, several witnesses were keen to stress that their concerns lay more with the AI capabilities than just the autonomous behaviour. Professor Noam Lubell, Professor at University of Essex School of Law, felt that existing definitions place too much emphasis on autonomy as a behaviour rather than the more concerning issue of the use of AI technologies to drive that behaviour.[64] This accords closely with our approach.

44. Professor Mariarosaria Taddeo, Associate Professor at the Oxford Internet Institute, told us:

> "We focus on the autonomy of these systems, but it is not just that. We have plenty of automatic weapons in place already, although some are not allowed, such as landmines. It is not just the autonomy that is a problem there. It is the learning ability and the adaptive behaviour of these systems that make it so problematic." [65]

45. Professor Dame Muffy Calder, Vice Principal and Head of College of Science and Engineering, University of Glasgow, and Specialist Adviser to this Committee, told us about the importance of identifying the function that the AI performs, asking " Where is the human agency in the system and where is the AI in the system? If the AI component is controlling air conditioning, is that such a concern? If it is controlling weapons firing, I imagine that it is more of a concern."[66]

Challenges in creating a single definition

46. The distinction between 'fully' and 'partially' AWS can be seen in the definitions put forward by several states to the Group of Governmental Experts. For example, France draws a distinction as follows:

- 'Fully' autonomous lethal weapon systems: Systems capable of acting without any form of human supervision or dependence on a command

[60] Mariarosaria Taddeo and Alexander Blanchard, Science and Engineering Ethics, *A Comparative Analysis of the Definitions of Autonomous Weapons Systems*

[61] Written evidence from Tsvetelina van Benthem (AIW0033)

[62] Q 144 (Tsvetelina van Benthem)

[63] Written evidence from Dr David Anderson (AIW0041)

[64] Q 1 (Professor Noam Lubell)

[65] Q 43 (Professor Mariarosaria Taddeo)

[66] Q 81 (Professor Dame Muffy Calder)

chain by setting its own objectives or by modifying, without any human validation, their initial programme or their mission framework.

- 'Partially' autonomous lethal weapon systems: Systems featuring decision-making autonomy in critical functions such as identification, classification, interception and engagement to which, after assessing the situation and under their responsibility, the military command can assign the computation and execution of tasks related to critical functions within a specific framework of action.[67]

47. We heard from many witnesses that the pace of technological innovation made producing a coherent, future-proofed definition challenging. Dr Elliot Winter, lecturer at Newcastle University Law School, noted that discussions at the Group of Governmental Experts had avoided discussions of definitions based on technological attributes, focusing instead on the extent of the link between machines and operators, an approach which he said is preferable as it is "flexible and future-proofed".[68] Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering at Palantir Technologies UK, suggested that "By focusing on the most operationally salient functions of AWS, and not on specific technology components, this definition will not readily succumb to technological obsolescence".[69]

The value of a definition

48. Much of the evidence we heard has been clear on the need to develop a single internationally agreed definition of AWS. Professor Taddeo told us that previous definitions provided by states have often been done so cynically, setting very high thresholds for what constitutes an autonomous system. This, she said, had created confusion, a sense that the topic was not being taken seriously, and a space "in which actors are free to design, develop and test these weapons without having to call them autonomous weapons systems".[70] She called for a "definition that is realistic, that is technologically and scientifically grounded, and on which we can find agreement in international fora to start thinking about how to regulate these weapons". [71] Professor Stuart Russell noted that without a common definition or greater specificity, discussions of a ban on AWS may result in states having differing conceptions of what is being considered.[72] For instance, many states use ambiguous terms such as "initial programme", "mission framework", "desired, "goals", "parameters" and "unpredictable".

49. Some contributors to our inquiry noted that the focus on producing a single internationally agreed definition may introduce unnecessary complexity with limited gain. Professor Toby Walsh, Chief Scientist at the AI Institute, UNSW Sydney, felt that "worrying about the definition of AWS is a distraction from making progress on regulating this space".[73] Charles Ovink, Political

[67] Convention on Prohibitions or Restrictions on the use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Non-exhaustive compilation of definitions and characterizations (10 March 2023), p 5:* https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_CRP.1_0.pdf [accessed 26 September 2023]

[68] Written evidence from Dr Elliot Winter (AIW0001)

[69] Written evidence from Courtney Bowman, Global Director of Privacy & Civil Liberties Engineering, Palantir Technologies UK (AIW0025)

[70] Q 45 (Professor Mariarosaria Taddeo)

[71] *Ibid.*

[72] Q 120

[73] Written evidence from Professor Toby Walsh (AIW0026)

Affairs Officer, UN Office for Disarmament Affairs, and Christopher King, Head of Weapons of Mass Destruction Branch, UN Office for Disarmament Affairs, both drew our attention to the lack of technical definition of a nuclear weapon, something they both felt had not had an effect on enforcement of the Nuclear Non-Proliferation Treaty (NPT).[74] Mr King noted that the lack of a definition in the Nuclear Non-Proliferation Treaty is "seen to be positive" and that any issues of compliance have been related to fissile material production.[75]

50. Dr James Johnson, Lecturer in Strategic Studies at the University of Aberdeen, disagreed, stating that the lack of a clear definition can lead to ambiguities and loopholes that "certain nations can exploit to develop their own nuclear technology that certainly skirt the edges of the treaty".[76] For example, the lack of a definition means that certain technologies (such as tactical nuclear weapons, AI-enhanced cyber weapons, or hypersonic and counterspace missiles) are not caught by the Nuclear Non-Proliferation Treaty.[77] As a result, he said that the lack of a definition makes "it even more challenging to monitor compliance and enforce the treaty."[78]

51. It could also be argued that nuclear weapons are not similar enough to AWS to provide a helpful comparison. Nuclear weapons are a specific military technology, whereas AI is a general-purpose technology, and the development of nuclear weapons requires enrichment of large amounts of nuclear material, whereas the software-based nature of AI makes it difficult to monitor and contain.[79]

52. We have heard extensive evidence on the difficulties of producing a single robust definition of what constitutes an AWS which captures only the systems of concern. Some witnesses thought that, under an excessively broad definition, an anti-personnel mine would be classed as an AWS.[80] Professor Lubell highlighted that it is difficult to produce a definition which does not include systems that have been in use for decades, such as active radar homing or high-speed anti-radiation missiles.[81] However, Georgia Hinds, Legal Adviser, International Committee of the Red Cross, stressed that producing a definition was only the first step in policymaking on AWS. She told us:

> "First, you capture all autonomous systems that may raise humanitarian concerns and that need to be regulated. That is based on the idea of them selecting and attacking targets, which is very different from non-autonomous systems. Within that, you then have the definition of certain types of autonomous weapon systems that might require specific prohibitions."[82]

[74] Q 19 (Charles Ovink) and Q 137 (Christopher King)

[75] Q 137 (Christopher King)

[76] Q 137 (Dr James Johnson)

[77] *Ibid.*

[78] *Ibid.*

[79] Yasmin Afina and Dr Patricia Lewis, Chatham House, 'The nuclear governance model won't work for AI' (28 June 2023): https://www.chathamhouse.org/2023/06/nuclear-governance-model-wont-work-ai [accessed 20 September 2023]

[80] Q 1 (Professor Noam Lubell)

[81] *Ibid.*

[82] Q 1 (Georgia Hinds)

53. **The UK's lack of an operational definition of AWS is a challenge to its ability to make meaningful policy on AWS and engage fully in discussions in international fora. Other states and organisations have adopted flexible, technology-agnostic definitions and we see no good reason why the UK cannot do the same.**

54. ***In acknowledgement that autonomy exists on a spectrum and can be present in certain critical functions and not others, the Government should without further delay adopt operational definitions of 'fully' and 'partially' autonomous weapon systems as follows:***

- ***'Fully' autonomous weapon systems: Systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator.***
- ***'Partially' autonomous weapon systems: Systems featuring varying degrees of decision-making autonomy in critical functions such as identification, classification, interception and engagement.***

Capabilities of AI systems

55. This section will consider the AI technology underpinning AWS, including issues of trustworthiness such as predictability and reliability, transparency and explainability; the robustness of datasets; and processes for testing, evaluation, verification and validation of systems.

56. As we saw in the previous section, much of the concern about AWS is focused on systems in which the autonomy is enabled by AI technologies, with an AI system undertaking machine analysis on information obtained from sensors to enable performance of the critical functions of acquiring, tracking, selecting, and attacking military objectives. The prevalent AI technology underpinning this machine analysis is machine learning (see Box 5). In considering the impact of AWS, as well as any potential regulation, it is important to understand the capabilities of the underlying technology.

Box 5: Machine Learning

Machine learning is a subfield of AI wherein computer systems are developed that can learn and adapt without following explicit human-prescribed representations of which aspects of the problem are important. Instead, algorithms and statistical models are used to analyse and draw inferences from patterns in data automatically. An algorithm or model is trained on a set of 'training data' to identify patterns or make predictions. Machine learning systems can be descriptive (they can explain what has happened), predictive (they predict what will happen), or prescriptive (they make suggestions about what action to take). Machine learning is currently the dominant subfield of AI.

Important categories of machine learning include:

- Supervised machine learning models, which are trained with labelled data sets. Data labelling involves adding one or more human labels to raw data, such as images, text files, or videos, to specify its relevance, facilitating accurate predictions. Models can categorise data according to the labels provided in the training data.

- Unsupervised machine learning models, which look for patterns in unlabelled data. Unsupervised machine learning can find patterns or trends that people are not explicitly looking for and may not be aware exist.

- Reinforcement machine learning models, which are trained through trial and error to take the best action based on a reward system. The model receives positive or negative rewards after each action, and the model learns to maximise the total reward.

- Generative models, such as ChatGPT, which use neural networks to identify the patterns and structures within existing data to generate new content.

Source: 'Machine learning, explained', MIT Sloan School of Management (2021): https://mitsloan.mit.edu/ideas-made-to-matter/machine-learning-explained [accessed 16 November 2023]. Nvidia, 'What is Generative AI?': https://www.nvidia.com/en-us/glossary/data-science/generative-ai/ [accessed 18 September 2023] and IBM, 'What is data labelling?': https://www.ibm.com/topics/data-labeling [accessed 16 November 2023].

Predictability and reliability

57. The Ministry of Defence is forthright about what it sees as the potential benefits of AI, with the Defence AI strategy stating:

> "AI is perhaps the most transformative, ubiquitous and disruptive new technology with huge potential to rewrite the rules of entire industries, drive substantial economic growth and transform all areas of society."[83]

58. However, it is also important to understand the risks in how machine learning systems operate, and how predictable and reliable they can be. Reliability is the characteristic of a system that will perform its intended function without failure in a given context for a given period of time. Predictability is the characteristic of a system wherein the actions and status of the system can be forecast sufficiently well within a given context.

59. Dr Vincent Boulanin, Director of Governance of Artificial Intelligence Programme, Stockholm International Peace Research Institute, noted the importance of reliability and predictability for commanders who, he said, "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".[84]

60. Dr Mikolaj Firlej, Lecturer in AI Law and Regulation, University of Surrey, said that, in his view, the biggest risk in machine learning systems is their "inherent unpredictability" warning that "The outputs of [machine learning] systems are only probabilistic. It means that [machine learning] systems

[83] MoD, *Defence Artificial Intelligence Strategy* (June 2022): https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy [accessed 11 October 2023]

[84] Q 21 (Vincent Boulanin)

can produce uncertain outputs rather than consistently producing the same results".[85] Professor Taddeo stressed that issues surrounding predictability are "intrinsic to the technology itself" and that it is "unfeasible to imagine that we could overcome them."[86] On the other hand, humans might also act on a probabilistic basis—it is not clear whether humans act any more predictably than a machine would, though it is important to consider what sort of categories of behaviour are predictable and acceptable in a given context. Issues of benchmarking humans against AI are discussed in Chapter 3.

61. Exact predictability may not always be required. An AI system will often be used because it can do something better than a human or create a novel solution that is beyond easy human imagination. In these instances, the AI system is typically useful because simply knowing generally how an AI will react appropriately—rather than its specific actions—may be sufficient.

62. Many critiques of AI-enabled weapon systems focus on their 'brittle' nature, for example the possibility of minor changes in the inputs making them unreliable, and their difficulty generalising beyond the bounds of the data used to train them. Minor changes in the inputs to an AI system, even those which are imperceptible to humans, can result in a substantial change in the output. One study showed a change in a single pixel is sufficient to cause machine vision systems to draw radically different conclusions about what they see.[87]

63. This is an issue which the Government acknowledged in its policy statement *Ambitious, safe, responsible*, saying:

> "The unpredictability of some AI systems, particularly when applied to new and challenging environments, increases the risks that unforeseen issues may arise with their use. The relative difficulties with interpreting how some forms of AI systems learn and make decisions present new challenges for the testing, evaluation and certification of such systems. In addition, the high potential impact of AI-enabled systems for Defence raises the stakes for potential side effects or unintended consequences, particularly when they could cause harms for those interacting with them."[88]

64. Many of our witnesses were deeply concerned about this unpredictability. Professor Noel Sharkey, Emeritus Professor of AI and Robotics and Professor of Public Engagement at University of Sheffield, raised concern about the ability of an AWS to operate in a complex situation such as the battlefield. He told the us "The trouble with the battlefield is that it is replete with unanticipated circumstances. When I say replete, possibly an infinite number of things can happen: a number of tricks, a number of spoofs—a number of different things."[89] This was echoed by evidence from Dr Ingvild Bode, Dr Hendrik Huelss, and Anna Nadibaidze, academics at the Center

[85] Written evidence from Dr Mikolaj Firlej (AIW0034)

[86] Q 43 (Professor Mariarosaria Taddeo)

[87] Jiawei Su, Danilo Vasconcellos Vargas, Sakurai Kouichi, Cornel University, 'One pixel attack for fooling deep neural networks' (17 October 2019): https://doi.org/10.48550/arXiv.1710.08864 [accessed 24 November 2023]

[88] MoD, '*Ambitious, safe, responsible our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence#using-ai-safely [accessed 27 September 2023]

[89] Q 113 (Professor Noel Sharkey)

for War Studies, University of Southern Denmark, who said that an AWS is highly likely to "encounter inputs for which it was not specifically trained or tested" reducing the human commander's or operator's ability to predict outcomes accurately.[90]

65. Professor Gopal Ramchurn, Professor of Artificial Intelligence, University of Southampton, told us that this inherent risk of unpredictable and unreliable systems must be considered in light of the operational environment. He said:

> " … these systems are probabilistic. They may not even correctly recognise the situation they are in in order to choose the right course of action. Even if they recognise the right situation, they might choose a random course of action that is unexpected. In those situations, what matters is the level of risk you incur. If it is a low-risk situation, you might be okay with a machine being automated and maybe making some mistakes that do not have a huge impact. In high-stakes and high-risk situations, you may want to have more control."[91]

66. Professor Ramchurn suggested that the level of required machine-operator interaction could be set in relation to this risk, saying "You can code the machine to ask for permission to take action; you can code the machine to act and then tell you what has happened; or you can ask the machine to do everything without telling you, depending on the level of risk it perceives."[92] This need to build in appropriate safeguards was echoed by Courtney Bowman who said that there was an opportunity to design mechanisms that allow systems to 'fail safely', "figuring out the appropriate moment for a human to step into the loop to provide the appropriate level of insight and accountability".[93]

67. Witnesses also raised concern about systems that continue to develop adaptively after deployment. While most AI systems are trained offline in advance using data sets, continual learning systems incrementally acquire, update, accumulate, and exploit knowledge throughout their lifetime.[94] While this can improve performance, it also reduces predictability and raises the danger that an AI system could be led astray, perhaps even deliberately, by new data. Dr Boulanin warned that AWS which had the ability to learn continuously, taking in new data and changing its parameters of use, would be problematic.[95] The difficulty of effective testing of AI-enabled AWS is discussed later in this Chapter and in Chapter 4.

Data sets

68. In addition to implementing appropriate human input and control in the design phase, having a sufficiently large, representative data set where any bias is transparent is critical to the success of a machine learning system. Dr Jurriaan van Diggelen, Senior Researcher in AI and Program Leader,

[90] Written evidence from Dr Ingvild Bode, Dr Hendrik Huelss and Anna Nadibaidze (AIW0015)

[91] Q 83 (Professor Gopal Ramchurn)

[92] *Ibid.*

[93] Q 26 (Courtney Bowman)

[94] Liyuan Wang, Xingxing Zhang, Hang Su, Jun Zhu, Cornell University, *'A Comprehensive Survey of Continual Learning: Theory, Method and Application'* (January 2023): https://arxiv.org/abs/2302.00487 [accessed 27 September 2023]

[95] Q 21 (Vincent Boulanin)

Human-machine Teaming, Netherlands Organisation for Applied Scientific Research, told us:

> "During development, we should make sure we have a diverse training set; try to anticipate in which kinds of situations this system will be used; and include examples from those situations in the training set… You should not just let your AI go out in the wild and collect any training data it encounters."[96]

69. James Cartlidge MP, Minister for Defence Procurement, told us that the Ministry of Defence uses both real and synthetic training data[97] "to provide a richer dataset for training".[98] The Integrated Review Refresh noted the need for the UK "to secure a leading role in data access and infrastructure, which will be critical to the UK's competitiveness when developing and using digital technologies such as AI."[99]

70. However, other witnesses were concerned about the availability of this data. Mr Ovink noted that datasets appropriate for military use are necessarily much smaller and more limited in scope than those used for civilian purposes, and that 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.[100] Dr Elke Schwarz, Reader in Political Theory at Queen Mary University London, agreed:

> "Systems that employ AI for the full kill-chain[101] are likely to be marred by incomplete, low-quality, incorrect or discrepant data. This, in turn, will lead to highly brittle systems and biased, harmful outcomes that will likely yield counterproductive outcomes. Autonomous systems tend to be built and tested on rather limited samples of data, sometimes synthetic data, sometimes inappropriate data—it is simply not possible to model the complexities of a battlefield accurately."[102]

71. Andrew Otterbacher, a Director at Scale AI and former Second Secretary at the US Department of State, noted that the problem is not necessarily only the quantity of data, but the possibility of "extracting actionable insights that can be derived from this data."[103] In order to be useable, he said that the data must meet several quality criteria: "it must be relevant to the policy questions at hand, accurate in its measurements or assessments, complete in its scope, consistent over time, and well-labeled for immediate ingestion into AI algorithms."[104] However, most incoming data "falls short" of these criteria.

[96] Q 85 and Q 88 (Dr Jurriaan van Diggelen)

[97] Synthetic data is made up, by humans or generated by machine, and may be representative of real-world data or meet certain conditions. One use of synthetic data is to stress test systems for difficult scenarios that might occur. As with real data, this may not always be accurate.

[98] Q 170 (James Cartlidge MP)

[99] HM Government, *Integrated Review Refresh 2023: Responding to a more contested and volatile world*, CP 811 (March 2023), p 57: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1145586/11857435_NS_IR_Refresh_2023_Supply_AllPages_Revision_7_WEB_PDF.pdf [accessed 22 November 2023]

[100] Q 17 (Charles Ovink)

[101] Definitions vary, but broadly this is the progression from identifying a target, dispatch of forces to the target, attack and destruction.

[102] Written evidence from Dr Elke Schwarz (AIW0009)

[103] Written evidence from Andrew Otterbacher (AIW0043)

[104] *Ibid.*

Instead, "incoming petabytes[105] are piled on top of previously collected exabytes[106] which in turn are piled on top of zettabytes[107]." He therefore proposed that the US, UK and other allies must "curate and exchange 'AI-ready' data that can be immediately used to extract meaningful insights."[108] The ability to interrogate how this data is generated is also important.

72. The Ministry of Defence has acknowledged the challenges associated with the availability of quality data, stating data can often be:

> " … badly curated, making it challenging, time consuming and cost intensive to access sufficient levels of machine-ready data to train AI models. Data ownership and the ability to share data can also present significant challenges; the MOD does not always own the data that it needs and there can sometimes be cultural, security and commercial challenges in sharing data more widely."[109]

73. The Ministry of Defence states that it is working to tackle these issues through implementation of the Defence Digital and Data Strategies, is examining options to make representative datasets available to suppliers and use appropriate synthetic data in less data rich environments, and maximising data-sharing opportunities with allies and partners.[110] Likewise, Paul Lincoln, Second Permanent Secretary at the Ministry of Defence, told us that the UK is sharing datasets with allies and partners.[111]

74. As part of the Data Strategy, the Ministry of Defence also acknowledges the need to ensure control over data, stating there will be an "assertive management of suppliers and contracts, working with Defence Commercial Function to ensure Defence retains sovereignty over its data" and that "Defence needs to be a better customer, having greater control within contracts, ensuring that Industry protects Defence's data and increases access to it."[112] It is also important to ensure proper security controls around the models that this data informs.

75. The use of biased datasets to train machine learning systems which may then go on to replicate and exacerbate those biases is a major concern amongst AI policymakers. In the Government's *Ambitious, safe, responsible* policy statement it lists 'bias and harm mitigation', including addressing bias in algorithmic decision-making, as one of its ethical principles for AI in defence.[113]

76. Some evidence we heard has highlighted the potential for issues of bias to apply to the use of AWS. Richard Moyes from Article 36 told us:

[105] 1,000,000 gigabytes

[106] 1,000,000,000 gigabytes

[107] 1,000,000,000,000 gigabytes

[108] Written evidence from Andrew Otterbacher (AIW0043)

[109] Written evidence from MoD (AIW0035)

[110] *Ibid.*

[111] Q 171 (Paul Lincoln)

[112] MoD, 'Data strategy for Defence' (27 September 2021): https://www.gov.uk/government/publications/data-strategy-for-defence/data-strategy-for-defence [accessed 4 September 2023]

[113] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]

> "If we start to develop target profiles based on datasets and algorithms, there is the potential for bias to come into that such that we start to identify, perhaps accidentally, perhaps deliberately, people of certain skin colour or age or gender characteristics as being combatants."[114]

77. Written evidence from the Women's International League for Peace and Freedom (WILPF) also highlighted the risk of gender bias in AI generally, and with the use of AWS specifically: "Facial recognition software struggles to recognise people of colour; voice recognition struggles to respond to women's voices or non-North American accents; photos of anyone standing in a kitchen are labelled as women."[115]

78. The Government's policy statement notes the need for an assessment and mitigation of potential bias, including through addressing bias in algorithmic decision-making, careful curation of datasets, setting safeguards and performance thresholds, managing environmental effects and applying strict development criteria for systems.[116] Dr Keith Dear, Managing Director, Centre for Cognitive and Advanced Technologies, Fujitsu, also highlighted the need to question whether appropriate procedural tools to assess data bias and embedded value judgements have been applied at the point data is cleaned and transformed. [117]

79. **In addition to implementing appropriate human input and control in the design phase, high-quality training data, where any bias can be identified and accounted for, is crucial to the development of robust AI models. However, real-world data to train AWS is limited in quantity and quality, and models and tools may be third party, in which case the training data and processes may not be available for inspection.**

80. ***We welcome the Government's commitment to ensuring the gathering and processing of high-quality data sets. In order to achieve this aim, the Government must dedicate sufficient resources to projects which further this goal, including the arrangement of data-sharing agreements with allied partners, and the continuous audit and independent certification of datasets as appropriate.***

Transparency and Explainability

81. It is often valuable to understand how a model operates, raising issues of transparency and explainability. Transparency in the context of machine learning models refers to any information about the design, development or operation of a model. It should be stressed that what sort of information will be useful to whom can vary significantly from one context to another. Typically, the purpose, structure and underpinning data of algorithms should be visible to the people who use and regulate those algorithms.

82. Many machine learning models are described as 'black boxes' wherein there is limited knowledge of the internal workings of a system. Tom Andrews, Founder and CEO, GCH Technologies Ltd, described certain machine learning models as "inherently opaque", writing that "We are often able to derive why it makes certain decisions, but even the engineers that designed it

[114] Q 112 (Richard Moyes)
[115] Written evidence from the Women's International League for Peace and Freedom (AIW0006)
[116] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence*
[117] Q 32 (Dr Keith Dear)

cannot reproduce the step-by-step decision as to why it chose to do A instead of B."[118] Alice Saltini, Research Coordinator for the European Leadership Network, said we remain "far from understanding the inner mechanics of AI systems and discern[ing] the basis for its decisions or predictions."[119] James Black, Assistant Director of the Defence and Security Research Group, RAND Europe, noted that the issue of how, precisely, decisions are made is not a new one, and that similar concerns about black box decision-making can be raised about human analysis. He told us that the use of computational models in decision-making forced those using those systems to come to "sharper, clearer and more binary" choices about risk. [120]

83. Many methods to improve algorithmic transparency focus on the concept of 'explainable AI', in other words methods to help clarify the reasoning behind decisions or predictions made by the model. This can include developing a model from the outset which is inherently interpretable[121] or producing visualisations of which features of an input most substantially affect the outputs of a model.[122]

84. Dr Adrian Weller, Director of Research in Machine Learning, University of Cambridge, and Specialist Adviser to this Committee, has noted that "Within machine learning, there is a general feeling that ""transparency"– like "fairness" – is important and good", but that while often beneficial, transparency is not a universal good.[123] There are different types of transparency, the utility of which may depend on context.[124] Professor Taddeo argued that transparency can come at the cost of efficiency and efficacy of an AI system, hence necessitating a trade-off.[125] Professor Calder warned that there is a risk that explaining how a system operates means "you have kind of given the game away to the adversary, who can then go and poison the data."[126] Research on how methods of explainability are used in practice also found that the majority of explainable AI deployments are not for end users affected by the model but rather for internal stakeholders for purposes such as model debugging.[127] Explainability can also be used as a means to an end, even potentially including a checkbox effort at fulfilling an organisation's stated commitment to transparency.[128]

[118] Written evidence from GCH Technologies Ltd (AIW0024). Note that Machine learning models are in a sense "fully transparent" to their developers in that they can see every step of computation by examining the instructions of the underlying computer code – however, knowing these low-level instructions may be insufficient to provide meaningful intelligibility about the principles underlying an algorithm's behaviour.

[119] Written evidence from Alice Saltini (AIW0023)

[120] Q 25 (James Black)

[121] Cynthia Rudin, 'Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead', *Nature Machine Learning*, vol. 1, (2019), pp 206–215: https://www.nature.com/articles/s42256–019–0048-x

[122] Aniek F Markus *et al*, 'The role of explainability in creating trustworthy artificial intelligence for health care: A comprehensive survey of the terminology, design choices, and evaluation strategies', *Journal of Biomedical Informatics*, vol. 113 (January 2021), p 4: https://www.sciencedirect.com/science/article/pii/S1532046420302835#s0035

[123] Adrian Weller, *Transparency: Motivations and Challenges* (August 2017), p 55: https://mlg.eng.cam.ac.uk/adrian/transparency.pdf [accessed 27 September 2023]

[124] *Ibid.*

[125] Q 59 (Professor Mariarosaria Taddeo)

[126] Q 88 (Professor Dame Muffy Calder)

[127] Bhat *et al*, *Explainable Machine Learning in Deployment* (10 July 2023): https://arxiv.org/pdf/1909.06342.pdf [accessed 27 September 2023]

[128] Bhat *et al*, *Explainable Machine Learning in Deployment* and Adrian Weller, *Transparency: Motivations and Challenges*, p 55

Testing, evaluation, verification and validation

85. Ensuring a system behaves as desired requires robust testing, evaluation, verification and validation processes—and still, a 100% guarantee of reliable good performance "in the wild" is often not possible. The draft articles submitted by the UK and other nations to the 2023 Group of Governmental Experts state that:

> "Autonomous weapon systems may only be developed such that their effects in attacks are capable of being anticipated and controlled as required in the circumstances of their use, by the principles of distinction and proportionality."[129]

86. Therefore, according to the UK position, a high level of predictability is required in order for AWS to be compliant with the principles of international humanitarian law (IHL) (discussed further in Chapter 4). In order to obtain this, effective testing of AWS is required. Professor William Boothby, Air Commodore (RAF Retired) and Honorary Professor at Australian National University, stressed the need to develop and maintain testing facilities and to work up measures of reliability to test against. He warned that the expertise to undertake that testing will be expensive to acquire and maintain.[130] Mr Bowman noted the need for AI systems to "prove themselves in the field" and be tested in near-live situations.[131] As Professor Calder said:

> "In general, for any system, until we deploy it in what we call the wild, in the real world, we just do not know. We can test, reason and analyse, but the proof of the pudding is in the eating. It is in the actual deployment, whatever the purpose of the system."[132]

87. However, "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."[133] While this applies to other forms of AI reliability-testing, such as autonomous vehicles, war presents especially high-stakes, confusing, and adversarial situations for AI.

88. Professor Ramchurn said there are currently no well-defined standards for the use of AI-based systems in AWS, highlighting an "urgent need to develop testing frameworks and standards that would help the defence industry build and deploy these weapons responsibly."[134] He noted that current frameworks used in the testing and verification of autonomous systems are not suitable for AI systems that work alongside humans, because the frameworks "cannot account for human behaviours and human unpredictability".[135]

89. The Ministry of Defence told us that it is examining its processes and compliance regimes to ensure they can meet the testing and assurance challenges posed by AI systems, "including by ensuring that key safety standards are defined, achieved and maintained via MOD Regulators while

[129] Convention on Prohibition or Restriction on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Draft articles on autonomous weapon systems - prohibitions and other regulatory measures on the basis of international humanitarian law*

[130] Written evidence from Professor William Boothby (AIW0003)

[131] Q 26 (Courtney Bowman)

[132] Q 85 (Professor Dame Muffy Calder)

[133] Q 15 (Charles Ovink)

[134] Q 93 (Professor Gopal Ramchurn)

[135] Q 85 (Professor Gopal Ramchurn)

providing agile alternative risk-based approaches to support operational requirements where required and appropriate."[136]

90. **Testing AWS properly against all possible scenarios which may arise after deployment is extremely challenging and indeed may be impossible. However, it is vital that only systems which meet sufficient, context-appropriate standards of reliability and predictability make their way into use.**

91. ***The Government must develop standards for use in the testing, verification and validation of autonomous weapon systems. These standards should cover but not be limited to aspects of data quality and sufficiency, human-machine interaction and appropriate transparency and resilience.***

Meaningful human control and accountability

92. Effective integration of humans and AI into weapon systems—human-machine teams—is essential to capitalise on the potential of AI[137] and to ensure that its use complies with international law, as discussed in Chapter 4. At the centre of this is ensuring 'meaningful human control'. Professor Christian Enemark, Professor of International Relations, University of Southampton suggested that the central question of whether control is 'meaningful' is "how should AI technology be used in a way that assists (or avoids disrupting) the proper exercise of human moral agency?"[138] Human moral agency is crucial because of AWS' lack of a moral sense on which to base decisions, with "no empathy or compassion, and no capacity to imagine or take responsibility for the consequences of their actions."[139]

93. The Government's position on fully autonomous lethal weapon systems is that it "oppose[s] the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifecycle" and that "there must be context-appropriate human involvement in weapons which identify, select and attack targets".[140] The Ministry of Defence has stated that different contextual factors may include "the purpose of use, physical and digital environment, nature of possible threats, risks associated with system behaviour, regulatory environment, and so on". Meanwhile, "human involvement" means "the various points throughout the system lifecycle at which authorised, suitably qualified and experienced people exercise judgement to influence, direct or limit the behaviour of an AI-enabled system and its effects". Such control, according to the Ministry of Defence, must be "exerted prior to, during and post-use, regardless of the AI capability." Examples of before use include, among other things, policy decisions, R&D activities, system design, risk management process, system test and evaluation, and training of operators. During use could include setting and updating of parameters and monitoring system performance, conduct of targeting activities, coordination of battlespace activities to achieve a military objective. The Ministry of Defence states that target

[136] Written evidence from MoD (AIW0035)

[137] MoD, *Human-Machine Teaming* (May 2018): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/709359/20180517-concepts_uk_human_machine_teaming_jcn_1_18.pdf [accessed 24 November 2023]

[138] Written evidence from Professor Christian Enemark (AIW0004)

[139] Written evidence from Drone Wars UK (AIW0008)

[140] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-Enabled capability in Defence*

clearance and target prosecution decisions require human decision-making (see Box 6).[141]

Box 6: Targeting Directives

For most modern Armed Forces, Targeting Directives are one of the principal mechanisms by which the application of lethal force on the battlefield is regulated and controlled. Specifically, a Targeting Directive will seek to ensure that lethal force is always used in compliance with IHL.

British military doctrine recognises three relevant levels at which military activity is planned and executed. The Strategic Level sets the overall objectives for military activity; the Operational Level translates those objectives into specific sets of military action; and the Tactical Level is the one at which individual military actions are executed.

The Targeting Directive is one of the control mechanisms which ensures that Tactical Level activity (for example engaging specific targets with lethal force) conforms with IHL in achieving the desired Strategic Level objectives. In doing so, the selection of targets, the levels or type of force to be used, and the risks of collateral damage or civilian casualties, are all taken into consideration in order that the military action conforms to all the principles of IHL.

A Targeting Directive draws its authority from a political level of approval. The dynamic nature of warfare and the context within which it is conducted will often mean, however, that a Targeting Directive contains a mixture of freedoms and constraints on military activity, particularly the use of lethal force. In practice this is likely to mean that some military activity is pre-authorised, some is delegated but subject to dynamic decision making at a specified level of authority, and some is retained at a political level of authorisation. The underpinning principle of delegation is that authority rests at the level which is best able to ensure that the requirements of IHL are met.

94. Professor Enemark proposed that control is meaningful if it:

(1) Involves performance of the system's 'critical' functions by a human. Critical functions are generally understood to include selecting and engaging targets.

(2) Is exercisable in a timely fashion. There must be the opportunity to override AI.

(3) Does not involve excessive trust in AI. Humans may experience 'automation bias' (the tendency for humans to depend excessively on automated systems) and overestimate of the accuracy and reliability of information provided by AI, or feel they have no choice but to trust the AI if a situation is fast-paced or complex.

(4) Enables accountability. AI is unblameable and unpunishable, so a human must be accountable.

[141] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

(5) Is a feature of the system's design. This could involve building in limitations on AI behaviour, such as limiting the speed of AI information-processing so that human operators are less likely to be overwhelmed.[142]

95. However, BAE Systems noted operational constraints on having human control: where communications links cannot be assured due to environmental factors or jamming; and in cases of 'machine-speed' warfare in which a human's ability to respond in sufficient time becomes a limiting factor.[143]

96. Control can happen on different levels. As put by Dr Boulanin, "there are a lot of different possibilities within the notion of control."[144] He noted that meaningful control could be achieved by:

- Implementing control in the design phase, by limiting the type of targets a system can engage, how it recognises a target or limiting its operation in terms of time or space;
- Implementing control over the environment, for example by deploying the weapon only in remote environments or putting up barriers or signs to prevent civilians accessing an area where a weapon may be in operation;[145] or,
- Implementing control through human-machine interaction, by ensuring that a human remains in a supervisory role and maintains the ability to intervene in the weapons' operation.[146]

97. The final method of control could operate in various ways. For instance, it could happen on a 'green-light' basis whereby there is a built-in presumption against engaging a target unless consented to by a human—or a 'red-light' basis where there is a built-in presumption in favour of proceeding unless stopped by a human.[147] Or, if designed to do so, the AI may communicate with the human when it is uncertain or where it encounters a situation it has not seen before. In these situations the human could step in, and over time would learn what the machine can and cannot do.[148] One challenge for this is that an AI system is likely to be updated frequently over time. It was suggested to us that the level of autonomy given to the AI should depend on the level of risk.[149]

98. In determining what level of human control is morally acceptable, there are two key contextual factors—the type of target and the environmental conditions. The type of target could be either material (such as incoming missiles) or humans. Morally, it is less serious to target inanimate objects so a lower degree of control may be ethically permissible. As part of this equation, the type of force should be considered—is it deadly or non-lethal? The environment also requires consideration—is it 'cluttered' with civilians or friendly personnel (such as a city), or 'low-clutter' (such as at sea)?[150]

[142] Written evidence from Professor Christian Enemark (AIW0004)
[143] Written evidence from BAE Systems (AIW0022), Q 81 (Professor Gopal Ramchurn) and Q 97 (Professor Durrant-Whyte)
[144] Q 20 (Dr Vincent Boulanin)
[145] We note the difficulty in achieving such an environment.
[146] Q 20 (Dr Vincent Boulanin)
[147] Written evidence from Professor Christian Enemark (AIW0004)
[148] Q 81 (Dr Jurriaan van Diggelen)
[149] Q 83 (Professor Gopal Ramchurn)
[150] Written evidence from Professor Christian Enemark (AIW0004)

Further complicating the picture, it is conceivable that, in the future, an AI system might enable more accurate and precise targeting, thereby reliably leading to less collateral damage.

99. Establishing control is essential in ensuring that AWS can be used effectively and in a way that is understood by operators and commanders. Courtney Bowman from Palantir told us that there are "fundamental limitations" on what machines can do in terms of distinction and proportionality calculations and that, if operators do not understand these limitations, then they could be challenged on their capacity to make sound decisions.[151] BAE Systems and James Black from RAND Europe also stressed the importance of training for operators and leaders, arguing that military decision-makers are not necessarily promoted based on their understanding of AI.[152]

100. This understanding is also required so that accountability for the actions of AI-enabled AWS can be properly traced. Professor Taddeo warned about possible scenarios where meaningful human control cannot be ensured, for instance "an officer being asked to take responsibility for something that they do not control, that they do not understand and that might put them in a horrifying situation."[153] She warned that full accountability and transparency may never be achieved with AI systems, and so there will always be a trade-off. She called for the development of thresholds and quantitative measures to allow those trade-offs to be made.[154] Professor Taddeo also stressed the need for overridability and kill switches in systems.[155]

101. Moreover, meaningful human control is central in compliance with international humanitarian law. Dr Boulanin told us that, for an attack by an AI-enabled system to be lawful, a human should be involved to determine that the attack is not prohibited under international law. The term "context-appropriate human involvement", he said, recognises that that involvement may depend on the characteristics of the system (its predictability and reliability) and those of the environment (how predictable it is and whether civilians or civilian objects are present).[156] IHL is discussed further in Chapter 4.

102. **Context-appropriate human control is a difficult concept to define, presenting challenges to the development of policy on AWS. Determining whether human control has been satisfied and setting a minimum level of human involvement in a system involves considering many nuanced factors such as the complexity and transparency of the system, the training of the operator, and physical factors such as when, where and for how long a system is deployed.**

103. ***We note the Ministry of Defence's definition of "context-appropriate" and "human involvement". The Government must ensure that human control is consistently embedded at all stages of a system's lifecycle, from design to deployment. This is particularly important for the selection and attacking of targets.***

[151] Q 32 (Courtney Bowman)
[152] Q 25 (James Black) and written evidence from BAE Systems (AIW0022)
[153] Q 59 (Professor Mariarosaria Taddeo)
[154] *Ibid.*
[155] Q 46 (Professor Mariarosaria Taddeo)
[156] Q 20 (Dr Vincent Boulanin)

104. ***The Government must ensure that any personnel required to use AWS have been provided with the training to ensure they have sufficient technical knowledge of how the system operates and its limitations, enabling operators to have confidence and capacity to override decisions where necessary. Such training needs to encompass the technical characteristics of systems, but also the exercise of human agency and legal compliance in controlling them.***

CHAPTER 3: AWS AND THE BATTLEFIELD OF THE FUTURE

105. Bringing AI onto the battlefield through the use of autonomous weapon systems could be a revolution in warfare technology and is one of the most controversial uses of AI today. We heard from some witnesses that Autonomous Weapon System (AWS) could be faster, more accurate and more resilient than existing weapon systems, could limit the casualties of war, and could protect "our people from harm by automating 'dirty and dangerous' tasks"[157] (see Box 7). Whether the UK is able to harness these benefits depends on how effectively the weapons can be adapted to the battlefield. As the US National Security Commission on Artificial Intelligence said, "Throughout history, the best adopters and integrators, rather than the best technologists, have reaped the military rewards of new technology."[158]

106. However, there are concerns about how AWS can be used safely and reliably, whether they risk escalating conflict more quickly, and their compliance with international humanitarian law (discussed further in Chapter 4) and the ethics of these systems (discussed further in Chapter 5).

107. Although a balance sheet of benefits and risks can be drawn, determining the net effect of AWS is difficult. This was acknowledged by the Ministry of Defence, which soberly noted "It is not possible to model the net effects that autonomous weapon systems may have on warfare in any meaningful way given the broad range of possibilities and future scenarios."[159]

Box 7: Potential battlefield benefits and risks of AWS

AI-enabled weapons, including AWS, may provide advantages in warfare, including:

- **Operational impacts**: AWS may have advantages including speed, agility, increased resilience and possible increased accuracy (for example, compared with manual or pre-set tasks), lack of fatigue and stress, and an ability to operate in inhospitable and remote environments.
- **Impacts on casualties**: AWS may remove combatants from the front line of combat, reducing risk. AWS would not attack out of retaliation, fear, or anger, would not make mistakes due to fatigue or stress, and may have improved accuracy, which could reduce civilian casualties.

However, they also pose operational risks:

- **Sufficiency of technology**: Whether AI technology can accurately identify and target threats is a fundamental question. Current AI systems are highly brittle and struggle to generalise or adapt to conditions outside of a narrow range of assumptions.

[157] Written evidence from MoD (AIW0035)

[158] National Security Commission on Artificial Intelligence, *Final Report* (1 March 2021), p 77: https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf [accessed 1 August 2023]

[159] Written evidence from MoD (AIW0035)

- **Escalation and proliferation:** Removing humans from the battlefield may reduce hesitancy to use force and thus escalate conflicts. The increased speed of autonomous systems, as well as any unintended behaviour, could risk inadvertent escalation and heighten crisis instability.
- **Accountability**: There is a lack of clarity about who, if anyone, is responsible for the actions of an autonomous system if it behaves unlawfully or not as intended.
- **Cyber security**: Making use of systems underpinned by computer software leaves them vulnerable to cyber-attack. Attackers could seek to take control of a system, disrupt operations, gather confidential information or tamper with the training data.

Source: RAND, 'The Risks of Autonomous Weapons Systems for Crisis Stability and Conflict Escalation in Future U.S.-Russia Confrontations' (3 June 2020): https://www.rand.org/blog/2020/06/the-risks-of-autonomous-weapons-systems-for-crisis.html [accessed 1 August 2023] and Christoph Bartneck, Christoph Lütge, Alan Wagner and Sean Welsh, An Introduction to Ethics in Robotics and AI (Cham: Springer, 2020), pp 93–100

108. This Chapter discusses the possible battlefield and strategic impacts of AWS, including the impact on the number and nature of casualties, its use by non-state actors, its impact on the speed of escalation, and AI's role in nuclear command, control and communications.

Changing paradigms of conflict

109. The new paradigm of warfare brought in by AI has been referred to as "algorithmic" or "mosaic",[160] while Chinese strategists have referred to it as "intelligentised" war.[161] Whatever it is labelled, the nature of warfare will increasingly be defined by the quality and use of data and software and how armies use AI-enabled weapons. Although this is not our focus, uses of AI that are not directly lethal have also made the battlefield, and the defence and security landscape more broadly, much more complex. For instance, a deepfake[162] video of Ukraine President Volodymyr Zelenskyy showed him calling on soldiers to surrender.[163]

110. This section will cover the impact of AWS on the number and nature of casualties. It will assess the impact both from the perspective of the use of AWS by the UK and its allies, as well as the use of AWS by enemy forces. It will also discuss the need for AI which is resilient to outside interference, including adversarial attacks and data poisoning attacks.

Casualties

111. We heard from some witnesses that AI has the potential to reduce casualties in war. Lord Sedwill, previously the National Security Adviser, argued that "we should not assume that AI will necessarily make all weapons more dangerous. In many cases, it will make them more precise. Therefore, we

[160] Bryan Clark, Dan Patt and Harrison Schramm, Center for Strategic and Budgetary Assessments, *Mosaic Warfare: Exploiting Artificial Intelligence and Autonomous Systems to Implement Decision-Centric Operations* (11 February 2020): https://csbaonline.org/uploads/documents/Mosaic_Warfare_Web.pdf [accessed 28 July 2023]

[161] Elsa Kania, Center for a New American Security, *Chinese Military Innovation in Artificial Intelligence* (7 June 2019), p 1: https://s3.us-east-1.amazonaws.com/files.cnas.org/backgrounds/documents/June-7-Hearing_Panel-1_Elsa-Kania_Chinese-Military-Innovation-in-Artificial-Intelligence.pdf [accessed 24 November 2023]

[162] The manipulation of facial appearance through deep generative methods.

[163] YouTube video, added by The Telegraph: https://www.youtube.com/watch?v=X17yrEV5sl4&ab_channel=TheTelegraph [accessed 9 October 2023]

should be able to minimise combat casualties, at least on our side, and civilian casualties on our adversary's side".[164] Dr Emma Breeze, Assistant Professor in International Criminal Law, University of Birmingham, noted that AI has the potential to improve situational awareness, reduce casualties, and mitigate mistakes.[165]

112. Others were more sceptical of these benefits. Dr Elke Schwarz, Reader in Political Theory at Queen Mary, University of London, thought that use of AI will shape human practices, outlooks and aims to such an extent that "AI warfare" will prevail over "human warfare", with the technical elements of war being prioritised over the human dimension.[166] She argued that "speed and efficiency (and thus lethality)" will be prioritised, raising the question of whether "accelerated wars conducted with AWS [can] be won by anyone?"[167] As expressed by the Under-Secretary for Multilateral Affairs and International Economic Relations for the Philippines, Carlos J. Sorreta, "AI in the military domain is ultimately about speed in waging war. Speed might be good for waging war, but perhaps not so much for peace. Delays in armed conflict are critical breathing spaces for diplomacy to work, for peace to be given a chance".[168]

113. Professor Russell pointed out that both sides may have access to such technology. He called the presumption that only one's own forces will have access the "sole-ownership fallacy". While this may have been the case for some technologies such as—for a time—Predator drones[169], "it neglects the possibility that one's opponents would have these weapons and … that non-state actors would also fire these weapons."[170] He noted that "If the other side has the weapons, our soldiers are exactly in the harm's way that we hope to put their soldiers in."[171] Indeed, use of autonomous weapons in Ukraine may be leading to higher casualties, in particular through drones dropping grenades and being used as spotters for artillery. As a result, traditional methods of protection, such as trenches, are becoming less viable.[172]

114. Either way, the exact impact of AI on the nature and number of casualties in war is unclear. This is partially due to the difficulty of benchmarking the ability of machines against humans. Benchmarking is the practice of evaluating performance by comparison with a standard and is fundamental to the decision to deploy an AI-enabled system, but may be problematic in practice.

115. Georgia Hinds from the International Committee of the Red Cross told us that comparisons between AI systems and human soldiers lack "empirical evidence" and that "Instead, we are engaging in hypotheticals where we compare a bad decision by a human operator against a hypothetically good outcome that results from a machine process."[173] Charles Ovink, Political Affairs Officer, United Nations Office for Disarmament Affairs, told the

[164] Q 101 (Lord Sedwill)

[165] Written evidence from Dr Emma Breeze (AIW0007)

[166] Written evidence from Dr Elke Schwarz (AIW0009)

[167] *Ibid.*

[168] *Ibid.*

[169] A remotely piloted aircraft that is employed primarily for intelligence-collection and secondarily for targeting of enemies.

[170] Q 123

[171] *Ibid.*

[172] *Ibid.*

[173] Q 4 (Georgia Hinds)

Committee that while “it has been argued that there could be improvements in accuracy or reductions in collateral damage, this has not yet been demonstrated”.[174] Francis Heritage, previously a Royal Navy Warfare Officer, said that “we have not even begun to benchmark”.[175] Benchmarking would require a range of tests (such as comparing false and true positive rates) in a range of different conditions (such as peacetime vs wartime) and environments (such as rural vs urban). As asked by Mr Heritage, “how can we even start to know if applying machine learning systems before the engagement has happened is itself ethical or technologically worthwhile? How would we know whether a misidentification by an algorithm would have reasonably also been made by a human?”[176]

116. When asked what benchmarking the Ministry of Defence undertakes in relation to targeting, Lieutenant General Tom Copinger-Symes, Deputy Commander, UK Strategic Command, Ministry of Defence, told us that comparisons between non-automated and automated systems occur “as part of our [the Ministry of Defence’s] value for money assessment”, although this is in relation to the use of AI in recognition rather than targeting.[177]

Sabotage and counter AWS

117. Beyond intrinsic issues, there is also the risk of interference which changes the behaviour of a system. Systems underpinned by computer software are vulnerable to cyber-attack. Attackers could seek to take control of a system, disrupt operations, gather confidential information or tamper with the training data. Interference goes both ways—while an enemy can tamper with one’s own or an allies’ equipment, it also enables countering of enemy AWS.

118. Professor Taddeo, Associate Professor, Oxford Internet Institute, told us that the risks associated with being the target of sabotage are increased when applied to AI, as “we have limited control over the effect that that [sabotage] may lead to”.[178] This leads to a situation in which “we have a very unwanted outcome and no culprit, because the sabotage was not intended to cause those outcomes, the scale and effects are too big and too wide, and we are not able to ascribe responsibility for those outcomes in a just and fair way.”[179] This is made more difficult by the range of parameters which may be tampered with.[180]

119. Dr Firlej, Lecturer in AI Law and Regulation, University of Surrey, argued that poisoning of machine learning data can lead to “the entire system” being “compromised”.[181] For instance, a Reaper drone[182] uses AI-augmented image recognition which collects data from the environment and recognises which targets should be engaged. If the drone is the subject of a data poisoning attack, whereby false data is injected, the drone may attack civilians or friendly forces. Dr Firlej argued that “The risk of misjudging a target can

[174] Q 15 (Charles Ovink)

[175] Written evidence from Francis Heritage (AIW0029)

[176] *Ibid.*

[177] Q 172 (Lieutenant General Tom Copinger-Symes)

[178] Q 49 (Professor Mariarosaria Taddeo)

[179] *Ibid.*

[180] Q 47 (Professor Mariarosaria Taddeo)

[181] Written evidence from Dr Mikolaj Firlej (AIW0034)

[182] A remotely piloted aircraft system used primarily for intelligence gathering and targeting, but also capable of close air support and strike missions.

be significant, irrespective of the fact that there is a human acting as a supervisor".[183]

120. To aid protection against sabotage, Professor Taddeo proposed testing AWS in parallel to deployment in the battlefield: "if we have a testing condition in which we can observe the same system doing the same operation, not in the wild but in an environment that we control, so as to know what those parameters should be as a baseline, and then compare those two things in order to see whether the second one is being manipulated."[184] While this may not allow instant intervention in the battlefield—if a weapon does not allow intervention after deployment—it would at least allow changes to the AWS ahead of its next iteration.[185]

121. James Black, Assistant Director of the Defence and Security Research Group, RAND Europe, spoke to the importance of using counter AWS technology against enemies to avoid a "race to the bottom".[186] Rather than deploying the same AWS as an enemy, he argued that the UK should develop "asymmetric responses, so that, when an adversary is employing AI-enabled systems or autonomous weapon systems in a way that we ourselves are not ethically comfortable in doing, it does not mean that we have to respond with tit for tat and lower our own ethical standards; it means that, instead, we have alternative capabilities that we can use to counter that."[187] He illustrated this using the example of "swarm" technology. If the UK were uncomfortable with using a technology, "we would really need to invest in counter-swarm technology such that others could not gain an advantage over us."[188]

122. **AI-enabled AWS could offer step changes in defence capability including increased speed, efficiency and accuracy. These capabilities, if realised, have the potential to change the nature of warfare and reduce casualties.** ***The Government must ensure that there is sufficient research and resources to realise this potential and it must be realistic about the capabilities and limitations of AI systems, benchmarking the performance of AWS against the operation and fallibility of non-AI-enabled and human-operated systems.***

123. ***We note with concern that at the moment there is not enough being done to protect UK systems from interference or attack, or to develop methods to counter the use of AWS by adversaries. It is one thing to deploy a system without challenge, but quite another to cope not only with enemy action but with the realities of the battlefield. The Government must recognise the risk posed to our own side by enemy AWS, avoiding a "sole-ownership fallacy",[189] and must take action to ensure the resilience, as far as possible, of the UK's own systems.***

[183] Written evidence from Dr Mikolaj Firlej (AIW0034)
[184] Q 47 (Professor Mariarosaria Taddeo)
[185] *Ibid.*
[186] Q 35 (James Black)
[187] *Ibid.*
[188] *Ibid.*
[189] That is, the presumption that only one's own forces will have access to a technology.

Use of AWS by non-state actors

124. This section discusses the use of AWS by non-state actors, including efforts necessary to reduce the risk of non-state actors obtaining AWS or producing AWS using commercially available AI systems and drones.

125. Proliferation of AWS increases the risk of these weapons falling into the hands of non-state actors. As noted by the Ministry of Defence, "proliferation of advanced AI solutions has potential to increase the threats from non-state actors either through direct technology transfers from hostile states or through repurposing commercial technologies."[190] Similarly, the Minister of State for Defence Procurement stated that "you have to work on the assumption that this could get into their hands and some of them will be working on it."[191] Mr Black of RAND said "We are starting to see the ability of these organisations to use robotics; we have seen ISIS using commercially available drones either for reconnaissance and targeting reasons, or as crude improvised explosive devices or delivery devices for them."[192] In 2017 Islamic State undertook drone attacks against the Peshmerga and French Special Forces in Northern Iraq. Since then, the US Department of Homeland Security has warned of terrorist groups applying "battlefield experience to pursue new technology and tactics, such as unmanned aerial systems".[193]

126. AWS that are cheap and commercially available—but typically rudimentary—are especially likely to be acquired.[194] The Ministry of Defence's Artificial Intelligence Strategy raises the potential security concerns from the development and use of AI weapon systems by other state and non-state actors. It says:

> "AI has potential to enhance both high-end military capabilities and simpler low-cost 'commercial' products available to a wide range of state and non-state actors. Adversaries are seeking to employ AI across the spectrum of military capabilities, including offensive and defensive cyber, remote and autonomous systems, situational awareness, mission planning and targeting, operational analysis and wargaming and for military decision support at tactical, operational and strategic levels. Adversary appetite for risk suggests they are likely to use AI in ways that we would consider unacceptable on legal, ethical or safety grounds."[195]

127. Professor Russell agreed that the low cost of some AWS increases their risk of use by non-state actors, comparing the proliferation of low-cost AWS to that of small-arms: "We know that there are in the order of 100 million AK-47s in non-government hands at the moment, so we would expect proliferation on that scale. These weapons would be very cheap."[196] He added that the impact is extenuated by the lack of requirement for human

[190] Written evidence from MoD (AIW0035)

[191] Q 185 (James Cartlidge MP)

[192] Q 27 (James Black)

[193] Alexander Blanchard and Jonathan Hall, Centre for Emerging Technology and Security, *Terrorism and Autonomous Weapon Systems: Future Threat or Science Fiction?*, p 1. Department of Homeland Security, 'National Terrorism Advisory System' (9 November 2017): https://www.dhs.gov/sites/default/files/ntas/alerts/17_1109_NTAS_Bulletin.pdf [accessed 1 August 2023]

[194] Alexander Blanchard and Jonathan Hall, Centre for Emerging Technology and Security, *Terrorism and Autonomous Weapon Systems: Future Threat or Science Fiction?*, p 1

[195] MoD, 'Defence Artificial Intelligence Strategy' (15 June 2022): https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy [accessed 1 August 2023]

[196] Q 123

intervention.[197] He argued that "because autonomous weapons do not need human intervention, by definition, one person or a small group of people can launch as many weapons as they can afford. If these were small grenade-carrying quadcopters or other kinds of kamikaze devices, you could be launching these weapons in the tens of thousands, or conceivably even in the millions."[198]

128. While agreeing that AWS in the hands of non-state actors are a threat, General Sir Chris Deverell, former Commander, Joint Forces Command, did not see the threat as existential. He accepted that the acquisition of less advanced AWS, such as loitering munitions, would be "very dangerous" and could "pose a threat to Downing Street", but questioned whether they would "pose an enormous threat to the existence of humanity".[199] To combat any threat that they present, the UK should think about how to counter them, he said.[200]

129. The risk from non-state actors is potentially increased by use of open-source software. We heard that, while open-source software (see Box 8) provides benefits in relation to testing and assurance of the quality of software,[201] its use could present security risks.

Box 8: Open-source software

Open-source software is software released under a licence in which the copyright holder grants users the right to use, change and distribute the source code to any person for any reason. Examples of well-known open-source software include Mozilla Firefox and VLC media player.

Most of the major AI models developed by large organisations, included those developed by OpenAI and Google, are closed-source or proprietary software (although OpenAI have announced an intention to develop an open-source model). Open-source AI models have been released, including open-source large language models (LLMs) which aim to compete with Open AI's GPT4 or Google's Bard. Critics have noted that many of these models draw heavily upon the outputs of large tech companies, either by reverse engineering closed-source models or by using the source code of Meta's leaked large language model, LLaMA. They have also highlighted that the substantial computing power needed to pre-train large language models may act as a barrier to the development of large language models by anybody other than the biggest tech companies.

[197] *Ibid.*

[198] *Ibid.*

[199] Q 154 (Sir Chris Deverell)

[200] *Ibid.*

[201] Q 103 (Professor Hugh Durrant-Whyte)

Open-source models were debated at the AI Safety Summit. The debate concluded that while these might pose risks for safety, they might also promote innovation and transparency. This may appear a reassuring conclusion, but while risks remain they demand vigilance.

Source: Synopsys, 'Open Source Software': https://www.synopsys.com/glossary/what-is-open-source-software.html [accessed 31 July 2023]. Insider, 'ChatGPT creator OpenAI is getting ready to release an open-source AI model, report says' (17 May 2023): https://www.businessinsider.com/openai-chatgpt-release-open-source-ai-model-2023-5 [accessed 1 August 2023]. Will Douglas Heaven, 'The open-source AI boom is built on Big Tech's handouts. How long will it last?', MIT Technology Review (12 May 2023): https://www.technologyreview.com/2023/05/12/1072950/open-source-ai-google-openai-eleuther-meta/ [accessed 1 August 2023]. Davide Castelvecchi, 'Open-source AI chatbots are booming—what does this mean for researchers?', Nature, vol. 618 (29 June 2023), pp 891–92: https://www.nature.com/articles/d41586–023-01970-6 [accessed 10 November 2023] and DSIT, FCDO, and Prime Minister's Office, ' Chair's Summary of the AI Safety Summit 2023, Bletchley Park' (2 November 2023): https://www.gov.uk/government/publications/ai-safety-summit-2023-chairs-statement-2-november/chairs-summary-of-the-ai-safety-summit-2023-bletchley-park [accessed 10 November 2023].

130. However, we heard that the ability of non-state actors to use software or to train models would be challenging. Professor Russell argued that it would be difficult for non-state actors to acquire the physical platforms that run software in large enough volume to present a threat.[202] Dr Keith Dear, Managing Director, Centre for Cognitive and Advanced Technologies, Fujitsu,[203] Yasmin Afina, Research Associate, Chatham House, and Sir Chris Deverell, concurred, saying that advanced software requires high computing power and hardware that only a "handful of companies" possess[204] and that it "would be hard for non-state actors to develop" AWS.[205]

131. Home construction of hardware could also prove challenging to non-state actors. While the dual-use nature of much autonomous technology means certain parts, such as the drone itself, would be commercially available,[206] these would have to be converted or cannibalised for parts. While online networks of drone hobbyists exist to provide advice on developing autonomous drones and doing so would cost "no more than a new smartphone", engineering the parts together and attaching a feasible payload requires technical knowledge.[207] Anything homemade would "lack both the robustness and the capabilities of more expensive military systems."[208]

132. To counter threats from non-state actors, Mr Black argued for multilateral safeguards. Mr Black applied his argument about avoiding a "race to the bottom" to non-state actors.[209] Rather, the UK should attempt to shape broader debate on AI governance while accepting that non-state actors "can defer from whatever normative and legal frameworks are agreed" while putting in place safeguards "by limiting their access to working with certain companies, buying certain products or services or working with certain states."[210] Rather than attempting to reduce the risk of non-state actors to zero, he argued that "We need to find those ways of learning to live with that

[202] Q 126
[203] Q 27 (Dr Keith Dear)
[204] Q 16 (Yasmin Afina)
[205] Q 154 (Sir Chris Deverell)
[206] Q 16 (Charles Ovink)
[207] Alexander Blanchard and Jonathan Hall, Centre for Emerging Technology and Security, *Terrorism and Autonomous Weapon Systems: Future Threat or Science Fiction?* (June 2023), p 2
[208] *Ibid.*
[209] Q 35 (James Black)
[210] *Ibid.*

as tolerably as possible, rather than thinking that we can eliminate that risk … we certainly cannot reduce it to zero."[211]

133. Dr Dear and Mr Black argued that existing deterrence theory does not map onto non-state actors.[212] Mr Black noted that deterrence theory has evolved out of Cold War nuclear deterrence and does not work when applied to the decentralised, loose, non-hierarchical command structures of non-state actors, "which do not lend themselves to influencing in the same way as a traditional top-down military adversary."[213] Dr Dear argued that "deterrence by denial" therefore tends to be the main strategy.[214] He used the example of barriers stopping cars coming onto pavements in Whitehall. He proposed that such a strategy would also be effective against AWS. For instance, low-cost drones can be countered with jamming software, air intercept, layered air defences and even chicken wire or a net.[215]

134. **The proliferation of commercially available drones, coupled with the widening availability of AI software, including open-source software, could enable non-state actors to produce AWS from widely available civilian technologies.**

135. ***The Government must demonstrate to Parliament that it is committed to ensuring 'deterrence by denial' to defend its own citizens from the use of AWS by non-state actors, as well as methods to limit the proliferation of the precursors of AWS.***

Arms control

136. One potential solution to preventing the acquisition of AWS by hostile state and non-state actors is through arms control. The established approach for enforcing compliance by states with weapons control regimes is through treaties establishing arms control mechanisms which usually depend on systems for verification and monitoring. No arms control regime exists for AWS, in the absence of agreement on how to regulate these weapons internationally—and there would be challenges in the development of any such regime.

137. Mr Otterbacher, Director at Scale AI, argued that, while "export controls are a necessary tool for mitigating risks associated with the global distribution of AI-related hardware, their effectiveness is a complex equation balanced between security concerns, innovation, and international cooperation". Although controls on hardware such as semiconductors, graphical processing units[216] and high bandwidth memory can slow the rate of development of AI systems, their effectiveness is largely determined by the will of the targeted country to build alternative supply chains or to replicate components domestically.[217]

138. Drawing on his experience with the Treaty on the Non-Proliferation of Nuclear Weapons, Dr James Johnson, Lecturer in Strategic Studies, University of Aberdeen, observed that ambiguities in the definition of what

[211] *Ibid.*
[212] Q 27 (Dr Keith Dear, James Black)
[213] Q 27 (James Black)
[214] Q 27 (Dr Keith Dear)
[215] *Ibid.*
[216] A device designed to accelerate computer graphics workloads.
[217] Written evidence from Andrew Otterbacher (AIW0043)

weapons are regulated can lead to loopholes in any verification regime.[218] His view was that a specific international body would need to be tasked with inspections and as technology advanced, these mechanisms would need to evolve. He also noted that any AWS arms control regime would need to strike a balance between regulation and technological progress. "Just as the Nuclear Non-Proliferation Treaty acknowledges the right to use nuclear technology for peaceful purposes, any treaty on lethal autonomous weapons should also recognise [the] beneficial uses of autonomous technology".[219]

139. Further challenges with applying a conventional arms control regime to AWS were highlighted by Professor Kenneth Payne, Professor of Strategy, King's College London, who noted that: "the signature for developing AI is quite small; you do not need those uranium enrichment facilities. A lot of it is dual use. You are talking about warehouses with computers and scientists. How can you monitor potential defection from any arms control regime?"[220] The software-based nature of AI also brings unique challenges, as highlighted by Dr Dear. Existing export controls and non-proliferation regimes focus on "old-school, traditional hardware such as missiles, engines or nuclear material". Software, conversely, is "a different proposition and, clearly, a challenge".[221]

Escalation and an AI arms race

140. This section sets out the evidence we heard on the notion of an AI arms race and the increased escalatory risk posed by autonomous weapons, and then considers the applicability of existing deterrence theory to AWS.

"Arms race afoot"?[222]

141. We heard differing views as to the extent to which there is an arms race around the development of AWS, and the utility of such a statement. Narratives around an arms race often focuses on 'the West' vs China. To some degree, this is true. Professor Jinghan Zeng, Professor of China and International Studies, Lancaster University, noted that China has been heavily investing in AWS in an attempt to become a "world-leading army". However, while a "major player", he noted that China is "still some way behind the US."[223]

142. Various witnesses noted that the situation is more complex than often portrayed. In the context of China, Professor Zeng suggested that "China is not a unitary actor".[224] Professor Payne told the Committee that he sees "an arms race afoot", but that, contrary to conventional wisdom that this is between the West and China, the AI weapons arms race is multipolar.[225] Professor Sir Lawrence Freedman, Emeritus Professor of War Studies, King's College London, also thought that arms races are more complex than sometimes portrayed, saying that there are two sorts: one where both sides are trying to do the same thing and one where one side is trying to work out how to neutralise the capability of the other. Sir Lawrence told us that: "It is not particularly useful just to think about everybody charging off in the

[218] Q 137 (Dr James Johnson)
[219] *Ibid.*
[220] Q 31 (Professor Kenneth Payne)
[221] Q 35 (James Black)
[222] Q 41 (Professor Kenneth Payne)
[223] Q 191
[224] Q 196
[225] Q 41 (Professor Kenneth Payne)

same direction and seeing who gets there first. It is a much more complex interaction with a variety of component parts, and the fundamental problem of offence and defence, of the interaction between the two opposing sides."[226] Dr Vincent Boulanin, Director of Governance of Artificial Intelligence Programme, Stockholm International Peace Research Institute, suggested that rather than an arms race, nations are engaged in an "AI capability race" with a race to ensure access to the building blocks of AI systems such as data, talent and hardware.[227]

143. Other witnesses highlighted the gap between the rhetoric of an arms race and the reality of a lack of action on AI development within the Ministry of Defence. We heard from Dr Dear about an instance where he was "stumped" when asked to provide an example of a major AI project within the Ministry of Defence.[228] Professor Payne echoed this, saying that there is a "phoney-war feeling" to the arms race discussion: "There is a disjuncture between what I am describing with thousand-strong aerial swarms and the reality of a small RAF experimental squadron firing its first missile from a drone." In contrast, he said that "It is much easier to see the radical nature of AI when you look at the basic research that companies such as OpenAI and DeepMind are doing than it is when you look at what is on the inventory of the RAF or the Army today."[229] However, the AI used by these companies (for text generation) is very different from the AI that would be useful in AWS (for image recognition).

144. Some witnesses noted that portraying an arms race as inevitable could lead to such a result. Mr Ovink noted that arms races are not inevitable, and that previous arms races and security dilemmas have been prevented.[230] Similarly, we heard from Dr Tom Watts, Leverhulme Early Career Researcher, Royal Holloway, University of London, that predicating policy on an imperative to develop because "competitors" and "adversaries" are doing so risks "generating a self-reinforcing cycle which could make the development of new types of autonomous weapon systems appear increasingly desirable."[231]

Stuck on the escalator?

145. We heard from many witnesses that AWS risk fuelling conflict escalation between states. Sir Lawrence Freedman noted that escalation can be both accidental—"you step on to the escalator and you cannot get off", as a result of misunderstanding the enemy's intention—or deliberate—where one side intentionally goes a rung up the "escalation ladder" in the hope of winning.[232]

146. Mr Ovink thought that AI applications in the military domain may lead to international instability, having the potential "to introduce elements of unpredictability at times of international tension. They can lead to actions that are difficult to attribute … This can create risks for misunderstanding and unintended escalation, which I think you can also agree is a serious concern."[233] Dr Ozlem Ulgen, Associate Professor in Law, University of Nottingham, expressed concerns about the lowering of thresholds for states

[226] Q 72
[227] Q 18 (Dr Vincent Boulanin)
[228] Q 41 (Dr Keith Dear)
[229] Q 33 (Professor Kenneth Payne)
[230] Q 18 (Charles Ovink)
[231] Written evidence from Dr Tom Watts (AIW0014)
[232] Q 74
[233] Q 15 (Charles Ovink)

to start wars and the resulting escalation of conflict. She argued that, while removing human combatants may reduce casualties among one's own force, overall civilian casualties may be increased. This "creates an unethical hierarchy of human dignity whereby those possessing the technology protect their own combatants from harm's way at the expense of and disregarding human targets."[234]

147. War games have shown that use of machines is likely to result in conflict escalating quicker than it would otherwise. First, machines may be more likely to escalate based on their assessment of risk, or 'escalation dominance'. This is only compounded by the greater speed with which the AI can make decisions when compared with their human counterparts. Second, humans competing against machines are more likely to escalate based on their predictions of how the enemy machine will act. Professor Payne pointed to a RAND US war-game which pitted two human-AI teams against each other where uncertainty about how much the adversary had outsourced to automatic decision-makers meant that they had to retaliate first, pushing up the escalation spiral.[235]

148. Mr Black noted that safeguards against escalation depend on understanding both human decision-makers and machine decision-making capabilities.[236] On top of this, it is important to understand the interaction between a human 'in the loop' and a machine.[237] He said that "The added challenge that we are encountering now is that we are not talking just about understanding human decision-makers in different national capitals around the world, how they will respond to our actions and how things may escalate; we are also trying to understand how their own approach to and integration of AI within their own decision-making will inform their escalation ladder and, therefore, how we can control moving up or down that escalation ladder."[238] Professor Payne placed a similar emphasis on understanding. He said that there was a knowledge gap in our understanding of the impact machine decision-making may have on escalation: " … deterrence, escalation and coercion are psychological as well as material factors. We have a decent understanding of how humans and human groups go about thinking about that. We do not have a similar level of understanding about how machines go about that."[239]

149. ***The development of AI capabilities, including AWS, has the potential to bring significant strategic benefits to the UK and its allies, for example enhanced conventional deterrence. However, the Government must not use AI-enabled AWS in a way that could result in unintended increases in escalatory risk.***

AI in nuclear command, control and communications

150. This section addresses the risks posed by the integration of AI into nuclear command, control and communications, also called NC3. It considers how the enhanced escalatory risk and uncertainty associated with AI systems intersects with the heightened risk inherent in nuclear command, control and communications.

[234] Written evidence from Dr Ozlem Ulgen (AIW0019)
[235] Q 24 (James Black)
[236] *Ibid.*
[237] 'Human in the loop' refers to an AI/ML system in which humans have a role in decision-making
[238] Q 24 (James Black)
[239] Q 24 (Professor Kenneth Payne)

151. Nuclear command, control and communications combines people, hardware (sensor, communications, and control technology) and software. The purpose of this combination is to enable commanders to target, operate, control, and use nuclear weapons by receiving data and advice from sensor systems and people tasked with interpreting it, to make decisions, and to send orders to nuclear forces to move, go on alert, or to strike targets.[240]

152. Use of nuclear weapons is primarily regulated by the Limited Test Ban Treaty (1963) and the Nuclear Non-Proliferation Treaty (1968), and the Comprehensive Test-Ban Treaty (1996),[241] all of which have been ratified by the UK.[242] However, these do not regulate Nuclear Command, Control and Communications. The Treaty on the Prohibition of Nuclear Weapons (2017), which comprehensively prohibits nuclear weapons, has not been signed by the UK or the other four nuclear weapon states parties to the Nuclear Non-Proliferation Treaty.[243] Some have also advocated creating domestic legislation regulating use of nuclear command, control and communications (see Box 9).

Box 9: Block Nuclear Launch by Autonomous Artificial Intelligence Act

In the US Congress, a cross-party group of Senators have introduced the Block Nuclear Launch by Autonomous Artificial Intelligence Act. This aims to codify the Department of Defense's policy in its 2022 Nuclear Posture Review to "maintain a human 'in the loop' for all actions critical to informing and executing decisions by the President to initiate and terminate nuclear weapon employment" in all cases. This follows the National Security Commission on Artificial Intelligence's recommendation that the US clearly and publicly affirms its policy that only human beings can authorise employment of nuclear weapons.

Source: Block Nuclear Launch by Autonomous Artificial Intelligence Act of 2023, US Department of Defense, 2022 National Defense Strategy of the United States of America, including the 2022 Nuclear Posture Review and the 2022 Missile Defense Review (October 2022): https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONAL-DEFENSE-STRATEGY-NPR-MDR.PDF [accessed 1 August 2023] and National Security Commission on Artificial Intelligence, Final Report (1 March 2021), p 10: https://assets.foleon.com/eu-west-2/uploads-7e3kk3/48187/nscai_full_report_digital.04d6b124173c.pdf [accessed 1 August 2023]

153. Advances in AI have the potential to have greater effect in nuclear command, control and communications. Machine learning could improve detection capabilities of early warning systems, improve the possibility for human analysts to do a cross-analysis of intelligence, surveillance, and reconnaissance (ISR) data, enhance the protection of the nuclear command, control and communications architecture against cyberattacks, and improve

[240] Peter Hayes, *Nuclear Command, Control and Communications (NC3) in Asia Pacific* (September 2021), p 5: https://cms.apln.network/wp-content/uploads/2021/09/Peter-Hayes_NC3_APLN-Special-Report.pdf [accessed 1 August 2023]

[241] The Russian Federation withdrew ratification of the Comprehensive Test Ban Treaty on 2 November 2023.

[242] UN, *Treaty banning nuclear weapon tests in the atmosphere, in outer space and under water* (October 1963): https://treaties.un.org/doc/Publication/UNTS/Volume%20480/v480.pdf [accessed 1 August 2023]. UN, *Treaty on the Non-Proliferation of Nuclear Weapons* (1968): https://disarmament.unoda.org/wmd/nuclear/npt/ [accessed 1 August 2023]. UN, *Comprehensive Nuclear-Test-Ban Treaty* (September 1996): https://www.ctbto.org/sites/default/files/Documents/CTBT_English_withCover.pdf [accessed 1 August 2023].

[243] UN, *Treaty on the Prohibition of Nuclear Weapons* (September 2017): https://treaties.un.org/doc/Treaties/2017/07/20170707%2003–42%20PM/Ch_XXVI_9.pdf [accessed 1 August 2023]. Other nuclear armed states include the USA, France, Russia and China.

the way resources, including human forces, are managed.[244] In addition, AI has the potential to automate simple and repetitive tasks, which are subject to human shortcomings and emotions, cognitive bias, and fatigue.[245]

154. In particular, there are opportunities in nuclear testing and planning: AI can predict effects of nuclear detonation and analyse large amounts of nuclear test data.[246] The Limited Test Ban Treaty banned real-life testing of nuclear weapons. Since then, the virtualisation of nuclear tests means that weapons scientists have employed lasers and supercomputers to understand nuclear weapons.[247] As noted by Lord Sedwill, the UK tests all warheads in a "virtual environment."[248] Similarly, AI can be used for strategic planning and war gaming, with AI-driven simulations helping military strategists analyse scenarios and predict outcomes, providing insights into an adversary's capabilities, intentions, and potential responses.[249]

155. Advanced testing of nuclear weapons using AI may also help balance communicating and safeguarding capabilities, to achieve deterrence. As Sir Anthony Finkelstein, former Chief Scientific Adviser for National Security, noted, "You want to ensure your actual systems and technology are not known, at least not technically, while ensuring that the presence of these assets is known."[250] AI could provide a demonstration of capability and a signalling of force that does not undermine the technology itself.[251]

Nuclear escalation

156. However, use of AI in nuclear command, control and communications also has the potential to spur arms races or increase the likelihood of states escalating to nuclear use—either intentionally or accidentally—during a crisis. AI does not have to be directly connected to nuclear launchers to be involved in this process. For instance, it could provide advice to humans on matters of escalation.[252] Mr Ovink noted that AI has the potential to aid decision-makers by allowing faster real-time analysis of systems and

[244] Vincent Boulanin, UNU-CPR Centre for Policy Research, 'AI and Global Governance: AI and Nuclear Weapons – Promise and Perils of AI for Nuclear Stability' (12 July 2018): https://unu.edu/cpr/blog-post/ai-global-governance-ai-and-nuclear-weapons-promise-and-perils-ai-nuclear-stability [accessed 27 September 2023]

[245] James Johnson, Modern War Institute at West Point, 'Rethinking Nuclear Deterrence in the Age of Artificial Intelligence' (28 January 2021): https://mwi.usma.edu/rethinking-nuclear-deterrence-in-the-age-of-artificial-intelligence/ [accessed 1 August 2023] and written evidence from Alice Saltini (AIW0023)

[246] Anna Heise, 'AI, WMD and Arms Control: The Case of Nuclear Testing', in Thomas Reinhold and Niklas Schöring (eds.), *Armament, Arms Control and Artificial Intelligence: The Janus-faced Nature of Machine Learning in the Military Realm* (Cham: Springer, 2022), pp 117–127

[247] Daniel Oberhaus, Wired, 'Nuclear Tests Have Changed, but They Never Really Stopped' (16 July 2020): https://www.wired.com/story/nuclear-tests-have-changed-but-they-never-really-stopped/ [accessed 1 August 2023]

[248] Q 107 (Lord Sedwill)

[249] Benjamin Jensen, Scott Cuomo and Chris White, War on the Rocks, 'Wargaming with Athena: How to make militaries smarter, faster, and more efficient with artificial intelligence' (5 June 2018): https://warontherocks.com/2018/06/wargaming-with-athena-how-to-make-militaries-smarter-faster-and-more-efficient-with-artificial-intelligence/ [accessed 1 August 2023] and written evidence from Alice Saltini (AIW0023)

[250] Alex Wilner, Casey Babb and Jessica Davis, Lawfare, 'Four Things to Consider on the Future of AI-enabled Deterrence' (25 July 2021): https://www.lawfaremedia.org/article/four-things-consider-future-ai-enabled-deterrence [accessed 9 August 2023]

[251] *Ibid.*

[252] Edward Geist and Andrew Lohn, RAND Corporation, 'How Might Artificial Intelligence Affect the Risk of Nuclear War?' (2018), p 2: https://www.rand.org/pubs/perspectives/PE296.html [accessed 1 August 2023]

data, and providing enhanced situational awareness. However, this may compress the decision-making timeframe and lead to increased tensions, miscommunication and misunderstanding, including between nuclear-armed states.[253] Christopher King, Head of Weapons of Mass Destruction Branch at the UN Office for Disarmament Affairs, stated that the use of AI in pre-delegation of the launch of nuclear weapons is "an extremely dangerous concept that could result in catastrophic outcomes". Rather than deterring enemies, "it ultimately increases the risks of accidental or misperceived nuclear use."[254] AI could also pose a potential threat to second-strike capabilities, potentially increasing the likelihood of a first strike in a "use it or lose it" scenario.[255]

157. The effect of AI on nuclear strategy also depends on adversaries' perceptions of its capabilities as well as on what it can actually do. For example, it is technically difficult for a state to develop the ability to locate and target all enemy nuclear-weapon launchers, so such an ability also creates a strategic advantage. States therefore seek this capability and might pursue it despite technical difficulties and the potential to alarm rivals and increase the likelihood of conflict.[256] In calculating deterrence, a range of circumstances therefore need to be considered: the impact of the actual capabilities, the perceived potential of those capabilities, and the premature use or fallibility of those capabilities.

158. AI could also affect the risk of nuclear 'close calls': incidents that might have led to an unintended nuclear detonation or explosion, but did not. These incidents typically involve a perceived imminent threat to a nuclear-armed state which could lead to retaliatory strikes against the perceived aggressor although they have also included mechanical or technical errors. The exact impact that AI would have is unclear. AI capability may not be robust enough to act better than a human would or on the other hand, if AI were effective, it could reduce the likelihood of human error and provide transparency. Nonetheless, as noted by James Baker, Executive Director (Policy and Operations), Labour for the Long Term, many close calls occurred as a result of false positive technical data. Conversely, accidents have been avoided due to questioning of data by human operators.[257]

Technological risks

159. The complexity and brittleness of AI presents risks. There is a risk that an adversary could hack the system, poison training data, or manipulate inputs.[258] Likewise, 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".[259] We heard from Alice Saltini, Research Coordinator, European Leadership Network, that "adding technical elements to nuclear decision-

[253] Q 15 and Q 134 (Dr James Johnson)

[254] Q 134 (Christopher King)

[255] Written evidence from Alice Saltini (AIW0023)

[256] Edward Geist and Andrew Lohn, RAND Corporation, 'How Might Artificial Intelligence Affect the Risk of Nuclear War?' (2018), p 1: https://www.rand.org/pubs/perspectives/PE296.html [accessed 1 August 2023]

[257] Written evidence from James Baker (AIW0031)

[258] Edward Geist and Andrew Lohn, RAND Corporation, 'How Might Artificial Intelligence Affect the Risk of Nuclear War?' (2018), p 2: https://www.rand.org/pubs/perspectives/PE296.html [accessed 1 August 2023]

[259] Q 134 (Christopher King)

making or decision support systems introduces a new source of errors, biases, and vulnerabilities that could remain hidden from operators."[260]

160. In particular, the latest generation of AI based on neural networks poses challenges. Neural networks use layers of artificial neurons to learn patterns and representations of data in matrices and form predictions or decisions based on what they have learned. However, AI can generate content that is interpreted by people as being factually correct, but it is not.[261] Ms Saltini stated that "This poses serious risks for AI integration in nuclear command, control and communications and adds to the brittleness of AI systems, which struggle with slight changes or deviations in data input that they have not been trained on".[262] She therefore questioned whether technology is ready "for integration with critical defense-related decisions" and proposed that a "moratorium on the use of AI in any critical element of nuclear decision making should be enacted until a greater understanding of how to build reliable AI systems and their internal functions is gleaned" and that integration of neural network in noncritical elements, while serving a purpose, "should be done with caution."[263] Dr James Johnson, Lecturer in Strategic Studies, University of Aberdeen, also argued for the need for a very high level of reliability. He argued that any systems 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." [264] He used the "buzzword" of "graceful degradation": 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."[265]

161. ***The risks inherent in current AI systems, combined with their enhanced escalatory risk, are of particular concern in the context of nuclear command, control and communications. The Government should lead international efforts to achieve a prohibition on the use of AI in nuclear command, control and communications.***

[260] Written evidence from Alice Saltini (AIW0023). Wilfred Wan, UNIDIR, *Nuclear Risk Reduction: A framework for analysis* (2019), pp 29–32: https://unidir.org/sites/default/files/2019–11/nuclear-risk-reduction-a-framework-for-analysis-en-.pdf [accessed 1 August 2023]

[261] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Wenliang Dai, Andrea Madotto and Pascale Fung, 'Survey of Hallucination in Natural Language Generation', *ACM Journals,* vol. 55 (March 2023), pp 1–38: https://dl.acm.org/doi/abs/10.1145/3571730 and written evidence from Alice Saltini (AIW0023)

[262] Written evidence from Alice Saltini (AIW0023)

[263] *Ibid.*

[264] Q 135 (Dr James Johnson)

[265] *Ibid.*

CHAPTER 4: INTERNATIONAL LAW

162. The United Kingdom is bound by obligations under international law with respect to the development and use of new weapons. This Chapter explores:

- whether AWS incorporating AI technology can operate in compliance with international law as it applies to the battlefield;
- how accountability and enforcement can be assured; and
- whether new international law is required.

AWS and international law

163. The use of Autonomous Weapon Systems (AWS) in armed conflict is primarily governed by international humanitarian law (IHL). IHL is the body of international law which regulates armed conflict between states and between states and non-state armed groups. IHL aims to protect civilians and other non-combatants (such as the wounded and sick, medical personnel and prisoners of war) and to prevent unnecessary suffering. The fundamental rules of IHL are longstanding and derive from both customary international law and international treaties, in particular the 1949 Geneva Conventions and their 1977 Additional Protocols.[266] Some treaties prohibit or impose restrictions on the use of specific types of weaponry.[267] These IHL treaties contain detailed provisions covering specific aspects of armed conflict, however there are four general principles which apply to the conduct of hostilities in all cases even where specific rules have not been prescribed. These are set out in Box 10.

Box 10: Basic Principles of international humanitarian law

Military necessity: Military necessity dictates that military force should only be used against the enemy to the extent necessary to achieve a legitimate purpose of the conflict.

Humanity: The principle of humanity forbids a party to a conflict from imposing any suffering, injury or destruction which is not necessary to achieve legitimate military purposes.

Distinction: Parties to a conflict must at all times distinguish between civilians and combatants and between civilian objects and military objectives. Attacks must be directed solely at combatants or military objectives and attacks that fail to distinguish between civilians and combatants are classified as indiscriminate and unlawful.

[266] International Committee of the Red Cross, 'International Humanitarian Law Databases': https://ihl-databases.icrc.org/en/ihl-treaties/treaties-and-states-parties [accessed 3 October 2023]

[267] See for example: Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, amended 2001. Convention on the Prohibition of the Use, Stockpiling, Production, and Transfer of Anti-Personnel Mines and on Their Destruction, 1997 and Convention on Cluster Munitions, 2008.

Proportionality: IHL does not prohibit attacks which may cause incidental harm to civilians or civilian objects, but attacks which cause disproportionate civilian harm relative to the military benefits are unlawful. Additional Protocol I to the Geneva Conventions defines a disproportionate attack as one that "may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."

Source: UK Joint Service Manual on the Law of Armed Conflict, JSP 383 (2004): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/27874/JSP3832004Edition.pdf [accessed 3 October 2023] and International Committee of the Red Cross, 'Protocols Additional to the Geneva Conventions of 12 August 1949', p 30: https://www.icrc.org/en/doc/assets/files/other/icrc_002_0321.pdf [accessed 3 October 2023].

164. As some witnesses noted,[268] other branches of international law may be relevant to the control or use of AWS, notably international human rights law. International human rights law is particularly relevant to the use of AWS in peacetime security contexts. Verity Coyle, Senior Adviser at Amnesty International, told us that there is a history of weapons developed for military use being adapted for use by police and other internal security forces.[269] Concerns raised about the use of AWS in warfare may therefore also become relevant to use outside armed conflict situations. However, the focus of this Report is to examine the development of AWS for use in a military context, so this Chapter will concentrate on the ability of AWS to comply with IHL as the primary law regulating the conduct of hostilities.

Compliance of AWS with IHL

165. Many witnesses expressed significant concerns about the ability of AWS to operate in compliance with IHL and in particular the principles of distinction and proportionality.[270] The application of these principles often requires difficult and subjective judgments to be made in the context of complex and rapidly evolving military scenarios. Such judgments are heavily dependent on context and the specific facts of the situation. Determining the proportionality of an attack requires a value judgment to be made about whether the civilian impacts are "excessive".

166. Some types of AWS have existed for years and have been assessed to be IHL-compliant. Dr Vincent Boulanin, Director of Governance of Artificial Intelligence Programme, Stockholm International Peace Research Institute gave some examples including the Phalanx[271] system, which has been in use since 1973. Phalanx is installed on a ship and can be turned to autonomous mode if there is a risk of an incoming attack. The system will automatically

[268] Q 43 (Verity Coyle), Q 148 (Tsvetelina Van Benthem) and written evidence from Professor Thompson Chengeta (AIW0020),

[269] Q 58 (Verity Coyle)

[270] QQ 1–14 (Georgia Hinds, Prof Noam Lubell, Dr Daragh Murray), QQ 43–63 (Verity Coyle, Prof Mariarosaria Taddeo, Dr Alexander Blanchard), QQ 109–119 (Richard Moyes, Prof Noel Sharkey, Dr Paddy Walker), Q121 (Prof Stuart Russell), Q 140 (Tsvetelina van Benthem, General Sir Chris Deverell), written evidence from Dr Elliot Winter (AIW0001), Dr Elisabeth Hoffberger-Pippan (AIW0002), Prof William Boothby (AIW0003), Women's International League for Peace and Freedom (AIW0006), Dr Emma Breeze (AIW0007), Drone Wars (AIW0008), Rebecca Hall (AIW0013), Dr Ingvild Bode, Dr Henrik Huelss, Anna Nadibaidze (AIW0015), Article 36 (AIW0017), Stop Killer Robots (AIW0018), Dr Ozlem Ulgen (AIW0019), Prof Toby Walsh (AIW0026), Prof Steven Haines (AIW0032) and UK Campaign to Stop Killer Robots (AIW0038),

[271] Raytheon, 'Phalanx Weapon System': https://www.raytheonmissilesanddefense.com/what-we-do/naval-warfare/ship-self-defense-weapons/phalanx-close-in-weapon-system [accessed 28 September 2023]

identify incoming threats via its target identification system and try to neutralise them, be they incoming missiles or aircraft.[272] The Phalanx system raises fewer concerns about IHL compliance because it operates in an environment (at sea) where the presence of civilians is usually less of a factor, the degree of control retained by the operator and the narrow parameters within which it functions.[273]

167. The AWS which give rise to concern from an IHL perspective are those which incorporate AI technology enabling the system to select and strike a target autonomously so that the operator is not in a position to make the judgements required by IHL. We heard from Georgia Hinds, Legal Adviser, International Committee of the Red Cross, that her organisation is most concerned from a legal and ethical perspective by AWS which self-initiate a strike once activated by a human. These systems react to environmental information detected through sensors and use a generalised target profile that has been input at the activation stage. "The difficulty … is that the user is not choosing and does not even absolutely know the specific target, the precise timing and the location of that force application."[274]

168. Several witnesses emphasised that, at least in the current state of the technology, machines are not capable of substituting for humans in making these assessments. Dr Elliot Winter, Lecturer at Newcastle University Law School, said:

> "While machines are adept at limited tasks such as classifying a weapon from visual imagery or winning board games such as chess or even the Chinese game 'Go', they do not possess the higher-level understanding and reasoning required to, for example, identify an injured or surrendering combatant or to make inherently impressionistic, non-formulaic, decisions about what level of collateral damage is tolerable for a given attack. Levels of artificial intelligence beyond what is currently available would be required for satisfactory judgement-making capacity; perhaps even 'artificial general intelligence' which is as intelligent as humans. According to software experts, such technology is unlikely to be available for at least 20 to 40 years, if at all. For now, only humans can make those decisions."[275]

169. Professor Noam Lubell, University of Essex Law School, also took the view that "at least for the time being" it was impossible to imagine a machine carrying out the balancing act required by the principle of proportionality, which judges civilian loss against military necessity. He noted that there are AI-based collateral damage estimation tools which can assess one side of the equation, but "balancing the two we leave to humans … The technology cannot do it so clearly it would be unlawful to use it."[276]

170. Richard Moyes, Managing Director, Article 36, highlighted the capacity of AI-based weapon systems to learn and adapt so that the operator cannot fully understand or predict how the system will function: "I do not think

[272] Q 19 (Dr Vincent Boulanin)

[273] Q 20 (Charles Ovink), Q 50 (Verity Coyle), written evidence from Dr Emma Breeze (AIW0007) and written evidence from ART-AI (AIW0016)

[274] Q 1 (Georgia Hinds)

[275] Written evidence from Dr Elliot Winter (AIW0001). We note that although recent generative AI models have begun to display general capabilities, there are different views about how long it will take to develop artificial general intelligence.

[276] Q 6 (Prof Noam Lubell)

that I am arguing for a complete removal of AI capabilities from all aspects of weapons systems. It is simply that the users of systems need to be able to sufficiently understand the system that they are using and what will trigger an application of force by that system in order to make reasonable determinations about the likely outcomes of using that system in a specific context."[277]

171. As we saw in Chapter 2, concerns have also been raised about the risk of bias in AI systems. Bias within a target profile gives rise to further potential IHL compliance concerns. Richard Moyes noted that in the US drone programme people killed in the vicinity of a drone strike have been assessed to be combatants rather than civilians if they are men between the ages of 16 and 70. "That does not align with the legal determinations as to whether people are targetable or not."[278]

Meaningful human control

172. In Chapter 2 we discussed how phrases such as "meaningful" or "context-appropriate" human control are frequently used to describe the degree of human intervention required to ensure that an AWS complies with IHL. The Ministry of Defence's 2022 *Ambitious, safe, responsible* policy paper[279] states that the government "oppose[s] the creation and use of systems that would operate without meaningful and context-appropriate human involvement throughout their lifecycle. The use of such weapons could not satisfy fundamental principles of international humanitarian law".[280]

173. Dr Boulanin explained that "meaningful human control" is designed to capture the idea that humans should keep agency over the decision to use force. "Meaningful" signifies that a merely supervisory human role may not be enough to ensure IHL compliance, notably because of problems 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."[281]

174. The difficulty, as Tsvetelina van Benthem highlighted,[282] is that such general phrases are capable of different interpretations. Does meaningful control mean intervention at the programming stage where you specify the parameters of action and you train your personnel? Or is it the requirement of explainability throughout the lifecycle of a system? Or is it a requirement to have direct human input before every individual use of force? Potentially divergent approaches are hidden within these general terms. This has particular relevance to the international debate on regulating AWS which we examine in the final section of this Chapter.

[277] Q 112 (Richard Moyes)

[278] *Ibid.*

[279] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-Enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]

[280] See also the letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[281] Q 20 (Dr Vincent Boulanin)

[282] Q 145 (Tsvetelina van Benthem)

IHL obligations during weapons development

175. Article 36 of Additional Protocol I to the Geneva Conventions imposes an obligation on states "in the study, development, acquisition or adoption of a new weapon, means or method of warfare … to determine whether its employment would, in some or all circumstances, be prohibited by [Additional Protocol I or other applicable international law]." This means that when the Government develops an AWS it must conduct an effective review of whether the AWS is capable of being used in a manner which is compliant with IHL. Article 36 requires an assessment of the legality of new weapons at "all stages of the weapons procurement process, in particular the initial stages of the research phase (conception, study), the development phase (development and testing of prototypes) and the acquisition phase (including 'off-the-shelf' procurement)."[283]

176. Witnesses noted deficiencies in the Article 36 process: only a small number of states have the capacity to conduct Article 36 reviews[284]; there is no international oversight of the review process and states are not required to disclose the results of their weapons reviews[285], nor is there any binding guidance on how reviews are to be conducted[286]. Nevertheless, the Ministry of Defence has emphasised the importance of the Article 36 process as one of the essential guardrails to ensure that AWS are compliant with IHL.[287] The Ministry of Defence provided detail on how they conduct Article 36 reviews and stressed that these safeguards mean that "a weapon incapable of complying with IHL will never enter the UK inventory for use in armed conflict."[288]

177. Testing new weapons during development is a key element of the Article 36 process. However, we were told that effective testing of AI-enabled AWS is problematic and may be impossible. Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley, said:

> "There are difficulties in testing for discrimination, but proportionality and necessity are things that are so context-specific and dependent on aspects of the overall military situation that it would be very difficult not only to design an AI system that could make that judgment reliably, but to develop any of kind of testing in the lab for those conditions. I am not sure how you would design situations that are fully representative of the kinds of situations that could occur in the field, where there are difficult judgments to make.[289]

178. Professor Noel Sharkey, Emeritus Professor of AI and Robotics and Professor of Public Engagement, University of Sheffield, was also sceptical about whether a testing process could guarantee compliance of such systems with IHL principles: "if you look at machine learning, … it is trained from examples, and you can give it billions of examples. Where the examples

[283] International Committee of the Red Cross, *Legal review of new weapons: Scope of the obligation and best practices* (October 2016): https://blogs.icrc.org/law-and-policy/2016/10/06/legal-review-new-weapons/ [accessed 28 September 2023]

[284] Q 113 (Prof Noel Sharkey)

[285] Q 22 (Dr Vincent Boulanin) and written evidence from Rebecca Hall (AIW0013)

[286] Written evidence from Rebecca Hall (AIW0013)

[287] Q 173 (Paul Lincoln) and written evidence from MoD (AIW0035)

[288] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[289] Q 121

for warfare are coming from I do not know. It will not be trained on the battlefield, because it can take millions of iterations to learn something and you do not want that many people to die while it is learning."[290] Moreover, the uncertainty and unpredictability of the battlefield is such that "You can test … all you want in simulation, but you will not know how [an AWS] will behave on the battlefield, particularly against other military systems." Dr Paddy Walker, Senior Research Fellow in Modern War Studies, University of Buckingham, agreed that training such weapon systems would "require unbelievably specific data training sets" which do not exist.[291] However, General Sir Chris Deverell, Former Commander, Joint Forces Command, was more sanguine about the possible development of rigorous Article 36-compliant testing for AI-enabled AWS.[292]

179. In the case of systems that continue learning, taking in new data and changing their parameters of use, legal issues may arise if the learning has affected the performance in a way that would impact on compliance with IHL.[293] Dr Alexander Blanchard, Digital Ethics Research Fellow, Alan Turing Institute, and Dr Boulanin told us that it may be necessary to adapt the way Article 36 reviews are conducted, including the need for additional reviews, to take account of the way in which AI systems can transform.[294] Yasmin Afina, Research Associate, Chatham House, supported the establishment of monitoring and auditing requirements to ensure that AI-enabled weapon systems have not changed in such a way so as to invalidate the results of the initial review. She noted that industrial secrecy posed challenges for independent monitoring but that ideally states would not "mark their own homework".[295]

180. Recognising these challenges, the Ministry of Defence told us that they:

> "are currently assessing whether the current approach to legal review of weapons requires adjustment for AI-enabled capabilities, working in collaboration with partners to identify international standards. One key point of consensus is the need for re-review if there is any material change to the performance data or the core considerations that comprise a review. The emerging consensus is that such reviews may be built into the operational maintenance of the system and conducted in a constant feedback manner to ensure rapid and agile processes that maintain legal standards."[296]

181. **We have heard significant concerns about the ability of AWS which use AI technology in the targeting process to be used in compliance with IHL. The Government also acknowledges that there must be "context-appropriate" human control over any AWS which can identify, select and attack targets.**

182. ***The Government must demonstrate that AI-enabled AWS which it develops or deploys will function under sufficient levels of human control to be compliant with IHL on the battlefield.***

[290] Q 113 (Prof Noel Sharkey)
[291] Q 113 (Prof Noel Sharkey and Dr Paddy Walker)
[292] Q 141 (General Sir Chris Deverell)
[293] Q 10 (Prof Noam Lubell) and Q 21 (Yasmin Afina)
[294] Q 59 (Dr Alexander Blanchard) and Q 21 (Dr Vincent Boulanin)
[295] Q 21 (Yasmin Afina)
[296] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

183. ***The Government must demonstrate to Parliament that it has in place an effective system to perform Article 36 weapons reviews for AI-enabled AWS, particularly AWS which continue to learn and modify their behaviour after they have been deployed, including setting thresholds for triggering a new review.***

Enforcement and accountability

184. There are different forms of accountability for violations of IHL.[297] States can be held responsible under international law. Additionally, individuals can be held accountable through military codes of justice and national or international criminal law if they are personally responsible for ordering or launching a military attack which does not comply with IHL. States are required to prevent and punish "grave breaches", the most serious violations of IHL, including by enacting legislation to prosecute individuals who have breached the rules.[298] At international level the International Criminal Court also has jurisdiction to prosecute individuals for war crimes where a state is "unable or unwilling" to do so.[299]

Individual accountability

185. Many witnesses emphasised that maintaining human control over the operation of AWS is essential to ensure accountability for their use.[300] In its written evidence the Ministry of Defence were clear that "Human responsibility and accountability for decisions on the use of weapons systems cannot be transferred to machines."[301] Georgia Hinds explained that IHL is about processes, not just results. An attack which results in excessive collateral damage may nevertheless be lawful if the commander went through the proportionality assessment in good faith and it is assessed that a reasonable commander would have made the same decision. "This is about human judgement and a reasoning process, which cannot be outsourced."[302] The importance of human judgement in relation to IHL is discussed in paragraph 172. The risks of anthropomorphising machines are discussed in paragraph 242.

186. Professor Lubell observed that, without general AI (which does not currently exist), AWS are "tools not agents"[303] like any other weapon system. He described a scenario whereby an investigation into an alleged violation of IHL involving an AWS would work back from the commander's decision-making into the mechanics and design of the system and, if necessary, into its algorithms. Ultimately, if investigators could not work out why the system had malfunctioned it would have to be taken out of use, otherwise the military decision-maker could be held accountable for knowingly deploying a potentially defective weapon.

[297] Q 143 (Tsvetelina van Benthem) and Q 12 (Prof Noam Lubell)

[298] International Committee of the Red Cross, 'Obligations in terms of penal repression': https://www.icrc.org/en/download/file/1067/obligations-in-terms-of-penal-repression-icrc-eng.pdf [accessed 3 October 2023]

[299] International Criminal Court, *Rome Statute of International Criminal Court*, Article 8 (war crimes jurisdiction) and Article 17 (admissibility): https://www.icc-cpi.int/sites/default/files/RS-Eng.pdf [accessed 3 October 2023]

[300] Q 12 (Prof Noam Lubell), Q 21 (Dr Vincent Boulanin), Q 113 (Richard Moyes), Q 140 (General Sir Chris Deverell), written evidence from Prof Steven Haines (AIW0032), Dr Ozlem Ulgen (AIW0019) and Rebecca Hall (AIW0013)

[301] Written evidence from MoD (AIW0035)

[302] Q 6 (Georgia Hinds)

[303] Q 12 (Prof Noam Lubell)

187. Other witnesses saw greater problems in ensuring accountability for the use of AWS, in particular in the case of machine learning technology.[304] Georgia Hinds noted that individual criminal liability in IHL often requires knowledge or intent to be established: "If you have a system that is producing its own results or continuing to learn, and it produces a result that is beyond human intent—that might not be a malfunction; it might be that the system views it as an optimisation—you have a fundamental break with individual criminal responsibility."[305] Dr Stephen Harwood, Department of Space and Climate Physics, University College London, made a similar point: "The 'programmer' who designs the AWS is distanced by the unpredictable learning capability of the system, with the analogy of whether parents can be held responsible for the actions of their children once they mature. Likewise, it is question[able] whether it is fair that the commanding officer who ordered the deployment is held accountable for a technology that can choose its own target."[306] Rebecca Hall also commented there may be an accountability gap if the acts of an AWS cannot be attributed to a human who can be prosecuted as a matter of international criminal law.[307]

188. **Human decision-making is central to legal accountability for the use of AWS. Accountability cannot be transferred to machines.**

189. ***The Government must commit to integrating meaningful human control into all AI-enabled AWS which it deploys so that human accountability can clearly be assigned for use of AWS on the battlefield.***

Does the regulation of AWS require new international law?

190. The international community has been debating the regulation of lethal AWS for several years. In 2016 a Group of Governmental Experts on Lethal Autonomous Weapons Systems was established under the Conventional Weapons Convention. This has been the primary forum for international discussion. In 2019 the states parties to the Conventional Weapons Convention endorsed a set of 11 Guiding Principles[308] drawn up by the Group of Governmental Experts which make some basic statements about the application of IHL to the development and use of lethal AWS, including an affirmation that IHL continues to apply in full and that "human responsibility for decisions on the use of weapon systems must be retained since accountability cannot be transferred to machines" (see Box 11). The Group of Governmental Experts has met regularly since 2019 but little further substantive progress has been made.

[304] Q 52 (Prof Mariarosaria Taddeo), Q 113 (Richard Moyes) and written evidence from Rebecca Hall (AIW0013)

[305] Q 12 (Georgia Hinds)

[306] Written evidence from Dr Stephen Harwood (AIW0010)

[307] Written evidence from Rebecca Hall (AIW0013)

[308] Annex III of the Final Report of the Meeting of the Contracting Parties to the CCW Convention (December 2019), p 10: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement [accessed 2 October 2023]

Box 11: Guiding Principles agreed at the 2019 Meeting of the Contracting Parties to the CCW Convention

In 2019 the states parties to the Conventional Weapons Convention endorsed a set of 11 Guiding Principles relating to the development and use of lethal AWS. These include:

- International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems;
- Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system;
- Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole;
- Accountability for developing, deploying and using any emerging weapons system in the framework of the [Conventional Weapons Convention] must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control;
- In accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made whether its employment would, in some or all circumstances, be prohibited by international law;
- When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered;
- Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems;
- Consideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations;
- In crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphised;

- Discussions and any potential policy measures taken within the context of the [Conventional Weapons Convention] should not hamper progress in or access to peaceful uses of intelligent autonomous technologies;
- The [Conventional Weapons Convention] offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems within the context of the objectives and purposes of the Convention, which seeks to strike a balance between military necessity and humanitarian considerations.

Source: Annex III of the Final Report of the Meeting of the Contracting Parties to the CCW Convention (December 2019), p 10: https://documents-dds-ny.un.org/doc/UNDOC/GEN/G19/343/64/PDF/G1934364.pdf?OpenElement [accessed 2 October 2023]

191. In July 2023 the United Kingdom held the monthly Presidency of the UN Security Council and convened a debate, chaired by the Foreign Secretary, on "Artificial Intelligence: opportunities and risks for international peace and security".[309] The Foreign Secretary did not refer to the regulation of lethal AWS in his statement, but the UN Secretary General took the opportunity of the debate to call for the prohibition of lethal AWS without human control. In parallel, the Secretary General published a policy paper *A New Agenda for Peace* which includes a recommendation to "conclude, by 2026, a legally binding instrument to prohibit lethal autonomous weapon systems that function without human control or oversight, and which cannot be used in compliance with international humanitarian law, and to regulate all other types of autonomous weapons systems".[310]

192. Research conducted by the Stop Killer Robots campaign indicates that 106 states have expressly supported a new legally binding instrument regulating lethal AWS. 10 states have opposed this while 53 others have not declared a position on the question.[311] The states listed as opposing a new legally binding instrument are Australia, Estonia, India, Israel, Japan, Poland, Republic of Korea, Russia, United States, and the United Kingdom. However, there are significant differences between the approach of the states in this list. Although a small minority, including Russia, appear to be trying to obstruct progress in the Group of Governmental Experts, others, including the United Kingdom, oppose a new treaty but support a different kind of outcome.

193. In November 2022, the United Kingdom joined 69 other states in endorsing a statement delivered by Austria at the UN General Assembly[312] illustrating that there is a fairly broad coalition of support for action on lethal AWS. The resolution stated that: "Going forward, we recognise the importance of focusing efforts in particular on elaborating the normative and operational framework regulating, where appropriate and necessary, autonomous weapons including through internationally agreed rules and limits." The UNGA statement also supported the two-pronged approach advocated

[309] Record of the meeting of the UN Security Council on 18th July 2023, S/PV.9381 https://documents-dds-ny.un.org/doc/UNDOC/PRO/N23/210/49/PDF/N2321049.pdf?OpenElement [accessed 4 October 2023]

[310] UN, *A New Agenda for Peace* (July 2023), p 27: https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-new-agenda-for-peace-en.pdf [accessed 28 September 2023]

[311] Automated Decision Research 'State Positions': https://automatedresearch.org/state-positions/ [accessed 23 November 2023]

[312] Joint Statement on Lethal Autonomous Weapons Systems First Committee, 77th UN General Assembly Thematic Debate – Conventional Weapons (21 October 2022): https://estatements.unmeetings.org/estatements/11.0010/20221021/A1jJ8bNfWGlL/KLw9WYcSnnAm_en.pdf [accessed 3 October 2023]

by the UN Secretary General and the International Committee of the Red Cross of prohibiting those lethal AWS which are incapable of being used in compliance with IHL and regulating others. That dual approach can also be found in the "Draft articles on autonomous weapon systems"[313] submitted by the UK and a small group of like-minded states to the Group of Governmental Experts meeting in May 2023 and the joint statement: "Translating Progress at GGE LAWS into a Substantive Outcome" which was delivered on behalf of 52 states at the same meeting.[314] Some differences between states seem therefore to be more of form than substance.

194. There was a similar difference of opinion in the evidence we heard on whether it is preferable to regulate AWS through a new legally binding treaty or alternatively through "soft law" measures.

195. Witnesses who supported the "soft law" approach made several arguments against a proposed new treaty. First, we heard concerns that arguing for new international law to regulate lethal AWS risks implying that there are gaps in the existing law which could weaken IHL protection.[315] Second, some witnesses argued that even if an international agreement could be reached, it was likely to employ vague and general terms such as "context-appropriate human control" that would not clarify states' IHL obligations in relation to AWS.[316] Tsvetelina van Benthem observed: "the discussion about human control places us in a position where we can agree on the need for it but fundamentally disagree on what we mean by it."[317] Third, several witnesses doubted that agreement on a new treaty would be possible in the current geopolitical environment.[318]

196. Witnesses on this side of the debate supported the Government's approach of seeking clarifications of existing IHL and developing best practice.[319] Professor Lubell noted[320] that this would not be a straightforward process and that a lot of detailed work would be needed to unpack how IHL applies in particular situations. Tsvetelina van Bentham told us that significant uncertainty exists in aspects of IHL relevant to AWS and that efforts should be made to clarify the law. Her view was that it is not necessary to specify how the law would apply in every situation "but we at least have to be clear about the elements of the different rules. … Unless we have some reasonable sense of what the law is, we cannot implement this" law in relation to AWS.[321]

[313] Convention on Prohibition or Restriction on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Draft articles on autonomous weapon systems - prohibitions and other regulatory measures on the basis of international humanitarian law (15 May 20223):* https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.4_US_Rev2.pdf [accessed 28 September 2023]

[314] Joint Statement: Translating the Progress at the GGE LAWS into a Substantive Outcome (15 May 2023): https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/230515_Draft_Joint_Statement_LAWS_GGE_May_delivered_on_behalf_of_52_states.pdf [accessed 16 November]

[315] Q 6 (Prof Noam Lubell), Q 22 (Yasmin Afina) and written evidence from Rebecca Hall (AIW0013)

[316] Q 7 (Prof Noam Lubell, Dr Daragh Murray)

[317] Q 140 (Tsvetelina van Benthem)

[318] Q 6 (Prof Noam Lubell), Q 22 (Yasmin Afina), Q 135 (Christopher King), Q 147 (Tsvetelina van Benthem), Q 148 (General Sir Chris Deverell) and written evidence from Dr Elliot Winter (AIW0001)

[319] Written evidence from MoD (AIW0035), Q 22 (Yasmin Afina) and Q 23 (Dr Vincent Boulanin)

[320] Q 6 (Prof Noam Lubell)

[321] Q 142 (Tsvetelina van Benthem)

197. Some witnesses who argued for clarification of existing IHL rather than the creation of new legal rules highlighted the possibility of developing a manual as an alternative "soft law" mechanism for providing guidance on how the law applies to AWS.[322] The Ministry of Defence has proposed such an initiative within the Group of Governmental Experts.[323] Manuals are guidance documents on the application of IHL to specific circumstances. They are generally created by groups of independent experts convened by like-minded states and are not legally binding. An example is the Tallinn Manual on cyber warfare which was commissioned by NATO.[324] However other witnesses doubted the usefulness of a manual on AWS on the grounds that existing manuals do not have official status nor global support.[325]

198. Advocates for a new legally binding treaty also made a range of arguments in support of their position. Georgia Hinds told us that the International Committee of the Red Cross recognised the risk, highlighted by Professor Lubell, of implying that the existing law is deficient. However, she said that in the case of AWS, it is clear from the Group of Governmental Experts' debate that states hold different views about the limits and requirements for the design and use of autonomous weapons that flow from existing IHL: "the fact that we do not have consensus on how the rules apply specifically to autonomous weapons is what makes it clear to us that we need something specific, shared and codified to provide that understanding and clarity."[326] She added that the International Committee of the Red Cross are wary of trying to clarify or interpret IHL principles within the Group of Governmental Experts because of the risk that this may lead to watering down existing law.

199. Several witnesses made the point that even if not all states sign up to an international treaty it has value in setting international norms which may in time evolve into customary international law.[327] Verity Coyle said that creating "a new legally binding instrument can set up systems for monitoring, information sharing and best practice, which can curb unlawful development and transfers."[328] Others noted the challenges of regulating nationally in the absence of a clear international framework.[329]

200. We heard that, given the lack of progress in international negotiations, it was likely that some states would take the initiative to move the discussions out of the Group of Governmental Experts (which operates under a consensus rule meaning that a small minority can hold things up) into the UN General Assembly or another international forum where decisions could be taken by majority vote.[330] There are precedents for moving negotiations out of the Conventional Weapons Convention framework in order to make progress without the consensus rule. This happened in the case of both the 1997

[322] Written evidence from Rebecca Hall (AIW0013), Prof Bill Boothby (AIW0003) and Q 147 (Tsvetelina van Bentham)

[323] Group of Government Experts on Lethal Autonomous Weapon Systems, 'United Kingdom Proposal for a GGE document on the application of international humanitarian law to emerging technologies in the area of lethal autonomous weapon systems (LAWS)' (March 2022): https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2022/gge/documents/UK_March2022.pdf [accessed 3 October 2023]

[324] Michael Schmitt (ed.), *Tallinn Manual on the International Law Applicable to Cyber Warfare* (Cambridge: Cambridge University Press, 2013)

[325] Q 121 and written evidence from Dr Elliot Winter (AIW0001)

[326] Q 7 (Georgia Hinds)

[327] *Ibid.*, Q 52 (Verity Coyle, Prof Mariarosario Taddeo) and QQ 114–115 (Richard Moyes)

[328] Q 52 (Verity Coyle)

[329] Q 8 (Georgia Hinds) and Q 118 (Richard Moyes)

[330] Q 116 (Richard Moyes)

Ottawa Convention on Anti-Personnel Landmines[331] and the 2008 Oslo Convention on Cluster Munitions.[332]

201. In October 2023, the First (Disarmament) Committee of the UN General Assembly considered lethal AWS as part of its debate on conventional weapons control. In her statement opening the debate, Mrs Mitzumi Nakamitsu, Under-Secretary-General and High Representative for Disarmament Affairs, urged states to act because the perils of weaponizing new and emerging technology "have perhaps never been as grave as they are now".[333]

202. The First Committee subsequently approved a draft resolution[334] calling on the Secretary General to prepare a report for the General Assembly session in 2024 setting out the views of states, international organisations, civil society, the scientific community and industry on "ways to address the challenges and concerns [of lethal autonomous weapon systems] from humanitarian, legal, security, technological and ethical perspectives and on the role of humans in the use of force".[335] It also proposes adding lethal AWS to the agenda of the 2024 General Assembly session. The draft resolution acknowledges the role of the Group of Governmental Experts, but it seems clear that the sponsors intend it to be a first step towards action on lethal AWS in the General Assembly if progress remains blocked elsewhere.

203. The UK voted in favour of the resolution but expressed optimism about progress in the Group of Governmental Experts.[336] We find this comment by the UK representative surprising given that, as the UN High Representative noted, discussion in the Group of Governmental Experts over the past 12 months has led to no substantive results. We agree with the UN High Representative's call for states to translate their commitment to address these emerging threats into action. The Government should step up and show leadership in these future discussions, consistent with their ambitions in the 2021 Integrated Review[337] to be more active in shaping the international order. It is not in UK interests "to leave international law and norm setting to others."[338]

[331] Ottawa Convention on Anti-Personnel Landmines, *Convention on the Prohibition of the Use Stockpiling, Production and Transfer of Anti-Personnel Mines and on their Destruction*: https://geneva-s3.unoda.org/static-unoda-site/pages/templates/anti-personnel-landmines-convention/APLC%2BEnglish.pdf [accessed 3 October 2023]

[332] Diplomatic Conference for the Adoption of a Convention on Cluster Munitions, *Convention on Cluster Munitions* (30 May 2008): https://www.clusterconvention.org/files/convention_text/Convention-ENG.pdf [accessed 28 September 2023]

[333] UN, '*With Peace and Security Architecture Imperilled, First Committee Must Spotlight Disarmament in Multilateral Efforts to Ease Tensions, Says High Representative* (2 October 2023): https://press.un.org/en/2023/gadis3709.doc.htm [accessed 14 November 2023]

[334] The resolution will now pass to the General Assembly plenary body for adoption. That vote will take place after the finalisation of this report but it is not expected that the outcome will be significantly different.

[335] UN, 'First Committee Approves New Resolution on Lethal Autonomous Weapons, as Speaker Warns—An Algorithm Must Not Be in Full Control of Decisions Involving Killing' (1 November 2023): https://press.un.org/en/2023/gadis3731.doc.htm [accessed 14 November 2023]

[336] *Ibid.*

[337] HM Government, *Global Britain in a Competitive Age: The Integrated Review of Security, Defence, Development and Foreign Policy,* CP 403, March 2021: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/975077/Global_Britain_in_a_Competitive_Age-_the_Integrated_Review_of_Security__Defence__Development_and_Foreign_Policy.pdf [accessed 28 September 2023]

[338] Q 116 (Dr Paddy Walker)

204. ***We call for a swift agreement of an effective international instrument on lethal AWS. It is crucial to develop an international consensus on what criteria should be met for a system to be compliant with IHL. Central to this is the retention of human moral agency. Non-compliant systems should be prohibited. Consistent with its ambitions to promote the safe and responsible development of AI around the world, the Government should be a leader in this effort.***

CHAPTER 5: UK DOMESTIC POLICY ON AWS

205. This chapter outlines how the Government has approached development and use of Autonomous Weapon Systems (AWS) on a domestic level, also drawing on the Government's broader position on AI development and regulation.

Translating strategy to operational practice

206. In *Defence Artificial Intelligence Strategy* and *Ambitious, safe and responsible*, the Government sets out its policies on AI weapons in defence.

207. In *Ambitious, safe and responsible* the Government raises the importance of "realising the benefits of AI" and "countering threats associated with the use of AI by others" and further states that this is one of the most "critical strategic challenges of our time".[339] It also lays out the five principles under which it will develop and deploy AI-enabled systems. They are:

- **Human Centricity:** consideration of the impact of any AI systems on humans throughout the lifecycle of the system.
- **Responsibility:** establishing human responsibility and accountability for AI-enabled systems.
- **Understanding:** ensuring that relevant individuals appropriately understand AI-enabled systems and their outputs.
- **Bias and harm mitigation:** requiring those responsible for AI-enabled systems to proactively mitigate risk and biases from the systems.
- **Reliability:** AI-enabled systems must be demonstrably reliable and secure.[340]

208. Various witnesses praised the Ministry of Defence for its willingness to create and publish such a strategy, with Dr Elke Schwarz, Reader in Political Theory, Queen Mary, University of London, calling it "laudable". Likewise, Dr Ingvild Bode, Dr Hendrick Huelss and Anna Nabidaidze, academics at the Center for War Studies, University of Southern Denmark, commended the UK for being one of the "few states that have publicly released their principles on the use and development of AI in the defence sector."[341]

209. However, we heard concerns about the documents lacking clear detail. Dr Schwarz argued that the five principles are really "core challenges" that the Ministry of Defence must consider when designing and implementing AI weapon systems and not necessarily "principles" from which they will be building a strategy.[342] This was a sentiment that Professor Taddeo agreed with, stating that having the principles "is a step in the right direction" but that it should be just the "first one".[343] Dr Blanchard agreed that the principles left "a lot to be done," and said that the next steps the Government

[339] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 27 September 2023]

[340] *Ibid.*

[341] Written evidence from Dr Ingvild Bode, Dr Hendrick Huelss, and Anna Nabidaidze (AIW0015)

[342] Written evidence from Dr Elke Schwarz (AIW0009)

[343] Q 63 (Professor Mariarosaria Taddeo)

needed to take was "bringing those principles down to that more granular application".[344] Similarly, Dr James Johnson told us that "the [Defence Artificial Intelligence] strategy report reads very much like an integrated report or review, rather than a strategy that lays out clearly defined choices and priorities," and added that the document was more of an "aspirational rather than an operational document."[345]

210. The Government has committed to promoting these values internationally. In *Defence Artificial Intelligence Strategy,* the Government states that it will promote "ethical approaches and influencing global norms and standards, in line with democratic values".[346] Likewise, Mr Lincoln, Second Permanent Secretary at the Ministry of Defence, told us that the UK is "working with Five Eyes partners and with other international fora … not only on international humanitarian law and how it applies to artificial intelligence but on ethical principles."[347] However, Richard Moyes, Managing Director at Article 36, said that he believes that the UK is not currently a leader in this area and "could do significantly more in international leadership on the issue." He expanded on this, telling the Committee that he "broadly agrees" with the UK's posture about its role, but he did not "see it in practice in terms of actually building partnerships and driving the conversation forward."[348] Dr Paddy Walker, Senior Research Fellow in Modern War Studies at the University of Buckingham, said that he believed the UK should understand that it is "not in our interest" to leave the norm-setting to others as that could lead nations such as Russia deciding on the scope and pace of regulation and responsible standards for AWS.[349]

211. As part of its effort to influence "global norms and standards", the UK, along with other nations, submitted a set of draft articles to the 2023 Group of Governmental Experts (discussed in Chapters 1 and 4).[350] Among other things, the draft articles state that AWS should not be designed to target civilians and conduct engagements that would not be the responsibility of the commanders and operators using the system.[351] The UK has not set out how it plans to operationalise these principles. However, drawing parallels with the US Department of Defense's Directive 3000.09 on 'Autonomy in Weapon Systems',[352] Mr Otterbacher proposed the following:

- Specialised training for service members and commanders to ensure comprehensive understanding and responsible usage of AI-enabled weapons. He said that "our experience shows Commanders need

[344] Q 63 (Dr Alexander Blanchard)

[345] Q 138 (Dr James Johnson)

[346] MoD, *Defence Artificial Intelligence Strategy* (15 June 2022), p 10: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 28 September 2023]

[347] Q 186 (Paul Lincoln)

[348] Q 116 (Richard Moyes)

[349] Q 116 (Dr Paddy Walker)

[350] Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects, *Draft articles on autonomous weapon systems - prohibitions and other regulatory measures on the basis of international humanitarian law* (15 May 2023): https://docs-library.unoda.org/Convention_on_Certain_Conventional_Weapons_-Group_of_Governmental_Experts_on_Lethal_Autonomous_Weapons_Systems_(2023)/CCW_GGE1_2023_WP.4_US_Rev2.pdf [accessed 28 September 2023]

[351] *Ibid.*

[352] Department of Defence, 'Directive 3000.09—Autonomy in Weapon Systems' (25 January 2023): https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf [accessed 20 September 2023]

training on how to implement AI into the decision-making workflow chain and should schedule multiple practical exercises beginning with Command and Staff tabletop exercises to full Field Training Exercises then incorporate AI enhanced decision-making tools and weapon systems."

- Development of Tactics, Techniques, and Procedures (TTPs) to ensure uniformity and efficacy in the use of AWS, including best practices for deployment, escalation of force, and decision-making hierarchies, among other operational considerations. These serve as a "practical guide for both commanders and operators in the field, helping to standardize procedures and mitigate risks".

- Integration of training modules and Tactics, Techniques and Procedures into a doctrinal framework, allowing "for a more structured approach to implementing AWS and provides a formal platform for periodic review and updating".[353]

212. ***The Government should make explicit how it intends to implement domestically the five principles outlined in Ambitious, safe and responsible and the draft articles submitted to the 2023 Group of Governmental Experts.***

213. ***The Government should set out its plans to become a leader in setting responsible standards at every stage of the lifecycle of AWS, including responsible development and governance of military AI. These standards should refer to the Ministry of Defence's Five Ethical Principles for AI in Defence.***

Procurement, innovation and talent

214. In its written evidence to us, the Ministry of Defence laid out its approach to procurement of AI. It stated, "We are also examining our processes and compliance regimes to ensure that we can meet the challenges of accelerating technological change." The Ministry of Defence also stressed that while it does not rule out incorporating AI within weapon systems, it is "not in the process of procuring Autonomous Weapons Systems". It explained that its approach will be one that enables the "adoption and exploitation" of AI systems across defence.[354]

215. As part of this, the Ministry of Defence has introduced the 'Commercial X' programme, which they state will "change procurement processes by focusing on digital solutions." Commercial X is intended to "bring new technologies to users faster" and speed up delivery, ensuring that front line forces have the technology that they need to meet changing requirements.[355] The Ministry of Defence noted that technology is developing fast and that "Defence needs to adopt commercial approaches that can exploit changing technology." In addition, the Ministry of Defence said that Commercial X "is working with suppliers to address entry barriers to MOD for small and medium size

[353] Written evidence from Andrew Otterbacher (AIW0043)

[354] Written Evidence from MoD (AIW0035)

[355] MoD, 'Ministry of Defence Commercial – Commercial X' (August 2023): https://www.gov.uk/guidance/ministry-of-defence-commercial-commercial-x [accessed 24 November 2023]

enterprises."[356] Mr Lincoln told us that Commercial X has resulted in a 50 per cent reduction in time for getting procurements out to contract.[357]

216. Despite this, in July 2023 the House of Commons Defence Committee published *It is broke — and it's time to fix it: The UK's defence procurement system*. The Committee "discovered a UK procurement system which is highly bureaucratic, overly stratified, far too ponderous, with an inconsistent approach to safety, very poor accountability and a culture which appears institutionally averse to individual responsibility."[358] Key recommendations included:

- The Ministry of Defence and Defence Equipment and Support should engage in more consistent dialogue with industry.
- The Ministry of Defence should put forward a plan on how it intends to help develop and foster the defence workforce over the next 10 years.
- The Ministry of Defence should make key trade-offs between capability, cost, time, and technical complexity much earlier in the procurement process when requirements are initially being set.
- The Ministry of Defence must develop a much greater sense of urgency in its procurement methodologies.
- The Ministry of Defence should take, if necessary, a more robust attitude towards its contractors if programmes get into serious difficulty.[359]

217. In oral evidence, we heard that many of these concerns are seen as well founded. James Black, Assistant Director of the Defence and Security Research Group, RAND Europe, said that defence "manifestly does not have perfect" industrial policy on acquisition and procurement[360] and Dr Keith Dear said that "unless something changes, I do not see how defence will keep up in this area."[361] Nicolas Jouan, Senior Analyst, RAND Europe, stated that while the creation of the Defence Artificial Intelligence Centre in the wake of the 2021 Integrated Review "bolstered technical expertise on AI and laid the groundwork for a more integrated approach to procurement within MOD", the source of the Ministry of Defence's procurement issues stem from "a variety of challenges deeply embedded within Ministry of Defence's acquisition process."[362]

218. The Ministry of Defence's procurement processes are particularly lacking in relation to software and data, both of which are important for the development and use of AI. Mr Jouan told us that "More is required to ensure the MoD's ability to acquire and exploit software and datasets for the eventual use of AI systems", pointing to the low levels of investment by the Ministry of Defence—this will "at best allow MOD to explore merely the possibilities of large-scale data exploitation but likely not to deliver next-generation capabilities."[363]

[356] *Ibid.*

[357] Q 167 (Paul Lincoln)

[358] Defence Committee, *It is broke - and it's time to fix it: The UK's defence procurement system* (Ninth Report, Session 2022–23, HC 1099), p 3

[359] *Ibid.*, pp 46–49

[360] Q 37 (James Black)

[361] Q 39 (Dr Keith Dear)

[362] Written evidence from Nicolas Jouan (AIW0040)

[363] *Ibid.*

219. Mr Black told us that Government and industry need to work closer on AI procurement, in a way that is "much more collaborative and less transactional than what we are used to in other areas of defence procurement", while the pace of development of AI is a "real challenge".[364] He noted that there are existing procurement processes for "traditional defence primes[365] in more traditional, well-understood areas of hardware and, to a lesser extent, software", even if these do not always work "terribly well."[366] Andrew Kinniburgh, Director-General, Make UK Defence, echoed the sentiment that procurement "is very skewed towards the big primes" at the cost of small and medium-sized enterprises.[367]

220. However, procurement structures do not exist for AI in defence where the market—rather than being one of "monopoly monopsony"–is one where the Ministry of Defence has to engage with multinational tech companies outside the UK.[368] Mr Kinniburgh argued that defence procurement is "hopelessly outdated for the world of AI".[369]

221. Mr Jouan echoed the need for the Ministry of Defence to forge long-term links with industry, especially to underpin the long-term development that goes into new AI systems. He said that "Such effort requires sustained, strong relationships between MOD and industry over long-term period with seamless transition between research and production."[370]

222. Dr Vincent Boulanin, Director of Governance of Artificial Intelligence Programme, Stockholm International Peace Research Institute, told the Committee that a similar issue exists in the innovation of new AI systems, commenting that "we see a trend where Governments are also considering approaching civilian companies because they are leading innovation in many areas." This is leading to a model where innovation is being driven by the "civilian sector and now the military sector is trying to find a way to adopt these civilian developments."[371]

223. Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley, pointed out that this raises issues surrounding the accountability and trustworthiness of these weapon systems. He said that while the private sector has "a lot of experience" ensuring that AI systems, such as large language models, behave themselves, these systems "continue to misbehave."[372] He added that often the "standard of quality in software is not as high as one would want for developing a weapon system" with software regularly needing to be updated.[373] Numerous witnesses noted that this is particularly concerning for AWS given the possible impact they may have if they work in ways that are unexpected or unpredictable.

224. Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering at Palantir Technologies UK, said that because of the requirement for the AWS to work as expected it is vital that the researchers,

[364] Q 29 (James Black)
[365] Large defence contractors.
[366] Q 37 (James Black)
[367] Q 201
[368] Q 37 (James Black)
[369] Q 201
[370] Written evidence from Nicolas Jouan (AIW0040)
[371] Q 17 (Dr Vincent Boulanin)
[372] Q 127
[373] Q 129

manufacturers, and programme developers maintain responsibility for not only "short-term testing and early deployment outcomes, but also long-term maintenance assurance and product liabilities for the sustained delivery of marketed results."[374] Mr Jouan argued that the procurement of AI systems is "fraught with grey areas of accountability", noting the lack of a definition of trustworthiness. He asked "what performance benchmark could help assess the reliability of systems that have never been used before?" and proposed that the Ministry of Defence use an AI auditing system to "decipher and keep in check the effect of AI in decision processes", with a pre-defined standard of performance that would allow the Ministry to "reconcile the actual added value of AI systems over traditional systems and keep manufacturers' accountability in check". Doing so would "require fundamental changes in the MOD's internal audit system".[375]

225. Another method of trying to increase accountability and reliability is through requiring manufacturers and developers to provide through-life support. However, this can be difficult for autonomous systems for two reasons:

- Maintenance of autonomous systems requires self-monitoring capabilities from platforms, which must rely on sensors and predictive maintenance algorithms to bring them back to repair facilities before they break down. The cost of such sensor capabilities and the availability of facilities specific to autonomous platforms has become an issue in the United States, where the integration of unmanned systems is more advanced than anywhere else (including the UK).[376]

- Regular software updating of autonomous systems. Mass data collected by sensors and real time swarm management of an unmanned fleet require sophisticated software in need of regular updates.

226. Consequently, Mr Jouan argued, Ministry of Defence acquisition officers must take these factors into account when establishing through-life support agreements. The involvement of developers should be "an integral feature with scheduled cyclical updates" with close involvement by Defence, Equipment and Support alongside Senior Responsible Owners.[377] Moreover, the Ministry of Defence should secure access to software maintenance services from the manufacturer and any potential subcontractors.[378]

227. Given the drive for innovation and the Government's reduced ability to shape this industry, witnesses pointed out that the Ministry of Defence will need to buy products off the shelf from private providers. Professor Sir Lawrence Freedman, Emeritus Professor of War Studies, King's College London, told us that the Ministry of Defence has traditionally "not been great" at buying equipment off the shelf, instead having "very specific military requirements". The challenge for the Ministry of Defence is "being very clear about your specifications … and then letting the contractor provide them rather than keep on changing all that."[379] Lord Sedwill, previously the National Security Adviser, stressed the importance of enabling systems to adapt to changes in hardware with minimal intervention from the user: "defence needs to get

[374] Written evidence from Palantir Technologies UK (AIW0025)
[375] Written evidence from Nicolas Jouan (AIW0040)
[376] *Ibid.*
[377] *Ibid.*
[378] *Ibid.*
[379] Q 78

much more sophisticated about ensuring that the providers of the hardware platforms are open to plug and play systems, not only AI software systems but other systems."[380]

228. James Cartlidge MP, Minister of State for Defence Procurement, said that the Government recognises that it needs to be "swifter and more agile because otherwise we will lose the competitive edge against our adversaries."[381] Mr Lincoln, Second Permanent Secretary, pointed to the Defence Command Paper, which commits to a procurement timeframe of "no more than five years for conventional platform-type technology, and when it comes to software we would set ourselves a timeframe of three years," as evidence of the Government's commitment to improving procurement.[382]

229. Mr Black noted that if procurement of AWS is going to be successful, the Ministry of Defence needs to build the skills that are required to achieve that goal, asking "Do we have the skills, knowledge and expertise within government at all levels to make informed decisions about what that industrial base looks like, who to work with, how to incentivise them, how to shape it, what kinds of products, outcomes and services we want to see".[383] Likewise, Sir Lawrence Freedman said that "It is vital to have people in government with sufficient authority and competence to be able to assess what they are getting properly, to ask the right questions, and to make sure that these legal and ethical, and political, questions are fed in and that you are not just giving contracts to people to do what they would have done anyway".[384]

230. Mr Jouan noted that the expertise required to scrutinise procurement offers depends on the "technological maturity of procurement offers themselves".[385] He distinguished between two sets of requirements: procurement of "existing mature AI systems" requires "rapid acquisition processes and doctrine adaptation" to take immediate advantage of the technology, whereas "development of defence-focused AI systems built from the ground up require long-term development and coordinated R&D between MOD and industry".[386] Mr Kinniburgh also stressed the importance of developing a "common language" between developers, manufacturers, and government.[387]

231. The need for more AI expertise within the Ministry of Defence was identified in its Defence Artificial Intelligence Strategy.[388] In written evidence to us, the Ministry of Defence said that it is developing a "Defence AI Skills Framework which will identify key skills requirements across defence." It said that this will be carried out by the Defence AI Centre, which will also be looking at developing the recruitment and retention offer for staff in a range of roles from skilled generalist to those with specialist skills."[389] The Minister for Defence Procurement told us that "We [the Ministry of Defence] are aware, albeit anecdotally, of a range of key factors that impact our ability to

[380] Q 108 (Lord Sedwill)

[381] Q 166 (James Cartlidge MP)

[382] Q 167 (Paul Lincoln)

[383] Q 37 (James Black)

[384] Q 77

[385] Written evidence from Nicolas Jouan (AIW0040)

[386] *Ibid.*

[387] Q 205

[388] MoD, *Defence Artificial Intelligence Strategy* (15 June 2022), p 20: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 20 September 2023]

[389] Written evidence from MoD (AIW0035)

develop, attract and retain SQEP[390] across AI disciplines, including pay and incentives (e.g. higher starting pay, golden handshakes, and referral bonuses), access to cutting edge IT and toolsets, and workload". However, "The MoD does not currently collect statistics for "AI" as an independent profession or job descriptor, nor is information systematically tracked relating to the reasons for which civilian and military personnel leave the Department."[391]

232. Witnesses were widely positive about the Government's steps. However, Sir Lawrence Freedman noted that if you are unable to pay a higher rate, "you will just not get the people" as some Civil Service jobs would require private sector entrants to "almost halve their pay." He also noted the importance of "creating an atmosphere of intellectual excellence, a feeling that you are doing something important".[392]

233. Courtney Bowman explained that one of Palantir's concerns related to "the Armed Forces' ability to access AI talent—that is, those with a background in fields such as computer science and data science, and underlying disciplines such as mathematics and physics."[393] Mr Bowman noted that "sharply uncompetitive remuneration is one reason for this, but there are others, including inflexible MOD career structures … and a wider lack of recognition of how those with AI skills can deploy their talent in service of a critical national mission."[394]

234. In oral evidence to us, the Minister recognised the issue of pay in bringing in and retaining high quality staff, stating that it is a "profound challenge". He said that "no matter what steps" are taken the Ministry of Defence will "never compete with the private sector potential that a person could earn." He did, however, point out that people would still be motivated to work at the Ministry of Defence by other factors such as patriotism and public service.[395] Mr Lincoln pointed to the new Government's Data and Digital Framework that is providing "approximately an extra 10% on peoples' salaries," and recognised that "we need to do our best in upskilling not just individuals but the workforce as a whole".[396] Lieutenant General Tom Copinger-Symes agreed with the Minister about retention of staff and added that the learning and development opportunities that were available are the best "chance to upskill in defence" and that this, alongside "the mission", retains people.[397]

235. **We heard widespread concern about the Ministry of Defence's procurement processes. While we appreciate the complexities, this is all the more concerning given the additional challenges of creating effective processes for AI in defence.**

236. ***The Government should set up an independent committee of experienced executives to overhaul its defence AI procurement system. The committee should in particular recommend the best way for the Government to specify objectives for systems in advance with***

[390] Suitably qualified and experienced personnel.

[391] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[392] Q 77

[393] Written evidence from Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering, Palantir Technologies UK (AIW0025)

[394] *Ibid.*

[395] Q 168 (James Cartlidge MP)

[396] Q 168 (Paul Lincoln)

[397] Q 168 (Lieutenant General Tom Copinger-Symes)

clear criteria and how these criteria should be continually monitored and enforced post-deployment, including regular independent AI auditing. As part of this, the Government should require that software developers and manufacturers provide effective through-life support to address any issues.

237. **Issues of pay and ethical concerns act as barriers to recruitment. AI is highly complex and requires a very high degree of knowledge and qualifications in order to develop it. This requires officials to be the "brightest and the best". But the Ministry of Defence is hamstrung by the Government's requirement that all staff should be paid using existing Civil Service paygrades. This has resulted in salaries offered by the Ministry of Defence being around 50 per cent of those offered by commercial enterprises. This situation cannot be allowed to continue.**

238. ***The Government must solve this problem. It must be able to deploy sufficient qualified staff to work on AI and to deliver demanding scrutiny of procurement offers from private developers and manufacturers. This might be achieved by establishing new pay scales, or by bringing in private sector staff on secondment. Either way, it will be challenging but absolutely necessary if we are to have the ability to compete on the international stage and safeguard our country.***

Ethics

239. The Ministry of Defence states that any use of AI to enhance defence processes, systems or military capabilities is governed by the AI Ethics Principles that it laid out in *Ambitious, safe and responsible*. It argues that "This is critical to retain the confidence of our people, our partners and our wider stakeholders including Parliament and the general public that Defence equipment is safe and reliable and would only be used responsibly in pursuit of legitimate military objectives."[398] It also declared that it believes that weapon systems that use AI "can and must be used lawfully, safely and ethically."[399] Paul Lincoln, Second Permanent Secretary, told us that the Ministry of Defence is implementing these principles as part of a joint services publication, which would require that developers ensure the principles are embedded within systems.[400]

240. Professor Eamonn O'Neill, Director at ART-AI, agreed with this in his written evidence, commenting that "AWS potentially offer more ethical use than traditional weapons," in certain circumstances. This could be particularly true in situations where AWS reduce the indiscriminate capacity of weapons, where Professor O'Neil suggests the use of AWS should be "encouraged."[401]

241. In *Ambitious, safe and responsible* the Government categorically states that "the United Kingdom does not possess fully autonomous weapon systems

[398] Written evidence from MoD (AIW0035)

[399] *Ibid.*

[400] Q 187 (Paul Lincoln)

[401] Written evidence from ART-AI (AIW0016)

and has no intention of developing them."[402] Many witnesses were glad that the Government has no intention of developing fully autonomous weapons, believing that their use would be unethical. Professor Christian Enemark, Professor of International Relations, University of Southampton endorsed this sentiment; however, he stressed the importance of not getting stuck on the issue as it risks "getting overtaken by disagreement about what autonomy, in general, ought to mean" and can obfuscate the conversation about how "even an increase in the number of weapon system functions performed by an AI might reduce human control of that system to a morally unacceptable level."[403] Witnesses generally agreed that the debate needs to be a matter of deciding where and how it is ethical to use AWS and where a human controller should be involved in AWS decision-making processes.

242. Dr Boulanin explained that this is particularly important as we should not "anthropomorphise autonomous weapons" and that we must understand that important ethical principles such as distinction and proportionality cannot be fully automated, meaning that the human in the loop is vital to ensuring that the UK uses AWS ethically.[404] Dr Payne suggested that the "ethics discussion" surrounding where and how AWS are used needs to take place "in society at large."[405] He explained that he favours "the establishment of some sort of democratically selected public commission," to work as a method of widening the public debate around the issue.[406] However, in his evidence Dr Keith Dear pointed out that this is not currently happening because the debate on AWS "continues to be led by companies," and not the public.[407]

243. There is no sign of a move towards democratisation of this debate. The Ministry of Defence does not "currently undertake monitoring or polling to understand public attitudes towards the use of autonomous weapons systems".[408] However, in 2022 the Ministry of Defence set up the AI Ethics Advisory Panel to act as a group of experts providing advice and scrutiny. The Panel is an informal advisory board reporting to the Second Permanent Secretary in his role as Senior Responsible Owner for AI Ethics in the department. Mr Lincoln, the current Second Permanent Secretary, told us that the Panel "involves combinations of defence, industry, academia and critics of government policy".[409] The Ministry of Defence describes the Panel as being "advisory only, and has no formal decision-making powers, but will be responsible for scrutinising the MOD's ongoing approach to responsible and ethical AI". The Department also notes that the Panel "has not been involved in the creation of policy related to Lethal Autonomous Weapons Systems, nor the department's policy on AI safety."[410]

[402] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 20 September 2023]

[403] Written evidence from Christian Enemark (AIW0004)

[404] Q 21 (Dr Vincent Boulanin)

[405] Q 36 (Professor Kenneth Payne)

[406] Q 42 (Professor Kenneth Payne)

[407] Q 26 (Dr Keith Dear)

[408] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[409] Q 188 (Paul Lincoln)

[410] MoD, *Ambitious, safe responsible: our approach to the delivery of AI-enabled capability in Defence*

244. The Ministry of Defence asserts that "[t]ransparency and challenge are central to our approach. We are currently exploring options to be more proactive in communicating the work of the EAP [Ethics Advisory Panel], including through publishing the Panel's Terms of Reference, membership, and meeting minutes and, potentially, through an annual transparency report."[411]

245. Professor Taddeo, a member of the Panel, told us: "A panel such as this is very much needed in any defence organisation because there is a tendency otherwise to flatten ethics on security or safety measures, or to devolve ethical responsibilities to practitioners, which might not have the required understanding to address the multiple trade-offs and balances that applying and thinking about ethical questions imposes."[412] She said that the advice of the Panel has been taken "very seriously."[413] However, she suggested that another "board should be put together to oversee or lead efforts on translating the principles into practice".[414] She added that this translation requires expertise that does not necessarily exist within the Ministry of Defence.[415]

246. As well as the ethical issues arising from how, when and whom to target, there are ethical consequences for the operator of a system. The remoteness of any operator that works with autonomous weapons (as well as operators involved in cyber and other forms of remote engagement) means that they may have a reduced connection to both their team and the individuals being targeted. As argued by David Wagner, US Air Force Officer, "automation further de-humanises combat, lowering the barriers to entry for war via the creation of psychological and physical distance between decision-makers and death."[416] Lieutenant General Tom Copinger-Symes acknowledged that while the military has approaches to "managing stress" "over there", "we need to work out how we bring it back here": "Some of that is about training, some of that is just about pastoral care and companionship, and some of that is also about leadership."[417] This is an important point not confined solely to AI, but nevertheless the Government needs to demonstrate that it is taking account of these issues in training and pastoral care.

247. ***The Government has asserted that transparency and challenge are central to its approach. From the evidence we have taken, we have not found this yet to be the case. The Government should increase the transparency of advice provided by the AI Ethics Advisory Panel by publishing its Terms of Reference, membership, agendas, and minutes, as well as an annual transparency report.***

248. ***The Government should immediately expand the remit of the AI Ethics Advisory Panel to review the practical application of ethical principles in armed conflict and to cover ethics in relation to the development and use of AI in AWS.***

[411] Letter from James Cartlidge MP to the Chair (13 November 2023): https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/3/correspondence/

[412] Q 61 (Professor Mariarosaria Taddeo)

[413] Q 62 (Professor Mariarosaria Taddeo)

[414] *Ibid.*

[415] *Ibid.*

[416] David Wagner, 'Lethal Autonomous Weapon Systems and the Ethics of Robo-Killing: The Potential Effects of Intelligent Weapons on 'Jus Ad Bellum' and 'Jus in Bello' Compliance by States in Future Conflicts', *The King's Student Law Review and Strife Journal*, Issue II (2019), p 67

[417] Q 180 (Lieutenant General Tom Copinger-Symes)

Democratic control

249. As we said at the start of this Report, the fast pace of development, as well as the lack of publicly available information on how AI is being developed, pose issues for Parliamentary scrutiny, democratic endorsement and public confidence. Parliament's capacity for oversight depends on transparency and availability of information, its ability to anticipate issues rather than reacting after the event, and its ability to hold ministers to account. To recapitulate our recommendations on this point:

250. ***The Government must allow sufficient space in the Parliamentary timetable and provide enough information for Parliament, including its select committees, to scrutinise its policy on AI effectively. We naturally understand that elements of policy development may be highly sensitive; but there are established ways of dealing with such information. Arguments of secrecy must not be used to sidestep accountability.***

251. ***The Government must ensure that it engages with the public on AI-enabled AWS. It must also ensure that ethics are at the heart of its policy.***

252. **Overall, we welcome the fact that the Government has recognised the role of responsible AI in its future defence capability. AI has the potential to provide key battlefield and strategic benefits. However, in doing so, the Government must embed ethical and legal principles at all stages of design, development and deployment. Technology should be used when advantageous, but not at unacceptable cost to the UK's moral principles.**

SUMMARY OF CONCLUSIONS AND RECOMMENDATIONS

1. The lack of available statistics on the UK's spending on AI in defence means that it is difficult to determine whether the level of spending is appropriate and to compare it internationally. (Paragraph 17)

2. *The Government must publish annual spending on AI in defence as part of the Ministry of Defence's Finance and Economics Statistics Bulletin series.* (Paragraph 17)

3. *The Bletchley Declaration of November 2023 is, inevitably, aspirational, but it is a start. We commend the contents of the Declaration and encourage the Government to apply its principles to AI in defence.* (Paragraph 24)

4. The UK's lack of an operational definition of AWS is a challenge to its ability to make meaningful policy on AWS and engage fully in discussions in international fora. Other states and organisations have adopted flexible, technology-agnostic definitions and we see no good reason why the UK cannot do the same. (Paragraph 53)

5. *In acknowledgement that autonomy exists on a spectrum and can be present in certain critical functions and not others, the Government should without further delay adopt operational definitions of 'fully' and 'partially' autonomous weapon systems as follows:*

   - *'Fully' autonomous weapon systems: Systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator.*

   - *'Partially' autonomous weapon systems: Systems featuring varying degrees of decision-making autonomy in critical functions such as identification, classification, interception and engagement.* (Paragraph 54)

6. In addition to implementing appropriate human input and control in the design phase, high-quality training data, where any bias can be identified and accounted for, is crucial to the development of robust AI models. However, real-world data to train AWS is limited in quantity and quality, and models and tools may be third party, in which case the training data and processes may not be available for inspection. (Paragraph 79)

7. *We welcome the Government's commitment to ensuring the gathering and processing of high-quality data sets. In order to achieve this aim, the Government must dedicate sufficient resources to projects which further this goal, including the arrangement of data-sharing agreements with allied partners, and the continuous audit and independent certification of datasets as appropriate.* (Paragraph 80)

8. Testing AWS properly against all possible scenarios which may arise after deployment is extremely challenging and indeed may be impossible. However, it is vital that only systems which meet sufficient, context-appropriate standards of reliability and predictability make their way into use. (Paragraph 90)

9. *The Government must develop standards for use in the testing, verification and validation of autonomous weapon systems. These standards should cover but not be limited to aspects of data quality and sufficiency, human-machine interaction and appropriate transparency and resilience.* (Paragraph 91)

10. Context-appropriate human control is a difficult concept to define, presenting challenges to the development of policy on AWS. Determining whether human control has been satisfied and setting a minimum level of human involvement in a system involves considering many nuanced factors such as the complexity and transparency of the system, the training of the operator, and physical factors such as when, where and for how long a system is deployed. (Paragraph 102)

11. *We note the Ministry of Defence's definition of "context-appropriate" and "human involvement". The Government must ensure that human control is consistently embedded at all stages of a system's lifecycle, from design to deployment. This is particularly important for the selection and attacking of targets.* (Paragraph 103)

12. *The Government must ensure that any personnel required to use AWS have been provided with the training to ensure they have sufficient technical knowledge of how the system operates and its limitations, enabling operators to have confidence and capacity to override decisions where necessary. Such training needs to encompass the technical characteristics of systems, but also the exercise of human agency and legal compliance in controlling them.* (Paragraph 104)

13. AI-enabled AWS could offer step changes in defence capability including increased speed, efficiency and accuracy. These capabilities, if realised, have the potential to change the nature of warfare and reduce casualties. (Paragraph 122)

14. *The Government must ensure that there is sufficient research and resources to realise this potential and it must be realistic about the capabilities and limitations of AI systems, benchmarking the performance of AWS against the operation and fallibility of non-AI-enabled and human-operated systems.* (Paragraph 122)

15. *We note with concern that at the moment there is not enough being done to protect UK systems from interference or attack, or to develop methods to counter the use of AWS by adversaries. It is one thing to deploy a system without challenge, but quite another to cope not only with enemy action but with the realities of the battlefield. The Government must recognise the risk posed to our own side by enemy AWS, avoiding a "sole-ownership fallacy", and must take action to ensure the resilience, as far as possible, of the UK's own systems.* (Paragraph 123)

16. The proliferation of commercially available drones, coupled with the widening availability of AI software, including open-source software, could enable non-state actors to produce AWS from widely available civilian technologies. (Paragraph 134)

17. *The Government must demonstrate to Parliament that it is committed to ensuring 'deterrence by denial' to defend its own citizens from the use of AWS by non-state actors, as well as methods to limit the proliferation of the precursors of AWS.* (Paragraph 135)

18. *The development of AI capabilities, including AWS, has the potential to bring significant strategic benefits to the UK and its allies, for example enhanced conventional deterrence. However, the Government must not use AI-enabled AWS in a way that could result in unintended increases in escalatory risk.* (Paragraph 149)

19. *The risks inherent in current AI systems, combined with their enhanced escalatory risk, are of particular concern in the context of nuclear command, control and communications. The Government should lead international efforts to achieve a*

prohibition on the use of AI in nuclear command, control and communications. (Paragraph 161)

20. We have heard significant concerns about the ability of AWS which use AI technology in the targeting process to be used in compliance with IHL. The Government also acknowledges that there must be "context-appropriate" human control over any AWS which can identify, select and attack targets (Paragraph 181)

21. *The Government must demonstrate that AI-enabled AWS which it develops or deploys will function under sufficient levels of human control to be compliant with IHL on the battlefield.* (Paragraph 182)

22. *The Government must demonstrate to Parliament that it has in place an effective system to perform Article 36 weapons reviews for AI-enabled AWS, particularly AWS which continue to learn and modify their behaviour after they have been deployed, including setting thresholds for triggering a new review.* (Paragraph 183)

23. Human decision-making is central to legal accountability for the use of AWS. Accountability cannot be transferred to machines. (Paragraph 188)

24. *The Government must commit to integrating meaningful human control into all AI-enabled AWS which it deploys so that human accountability can clearly be assigned for use of AWS on the battlefield.* (Paragraph 189)

25. *We call for a swift agreement of an effective international instrument on lethal AWS. It is crucial to develop an international consensus on what criteria should be met for a system to be compliant with IHL. Central to this is the retention of human moral agency. Non-compliant systems should be prohibited. Consistent with its ambitions to promote the safe and responsible development of AI around the world, the Government should be a leader in this effort.* (Paragraph 204)

26. *The Government should make explicit how it intends to implement domestically the five principles outlined in Ambitious, safe and responsible and the draft articles submitted to the 2023 Group of Governmental Experts.* (Paragraph 212)

27. *The Government should set out its plans to become a leader in setting responsible standards at every stage of the lifecycle of AWS, including responsible development and governance of military AI. These standards should refer to the Ministry of Defence's Five Ethical Principles for AI in Defence.* (Paragraph 213)

28. We heard widespread concern about the Ministry of Defence's procurement processes. While we appreciate the complexities, this is all the more concerning given the additional challenges of creating effective processes for AI in defence. (Paragraph 235)

29. *The Government should set up an independent committee of experienced executives to overhaul its defence AI procurement system. The committee should in particular recommend the best way for the Government to specify objectives for systems in advance with clear criteria and how these criteria should be continually monitored and enforced post-deployment, including regular independent AI auditing. As part of this, the Government should require that software developers and manufacturers provide effective through-life support to address any issues.* (Paragraph 236)

30. Issues of pay and ethical concerns act as barriers to recruitment. AI is highly complex and requires a very high degree of knowledge and qualifications in order to develop it. This requires officials to be the "brightest and the best".

But the Ministry of Defence is hamstrung by the Government's requirement that all staff should be paid using existing Civil Service paygrades. This has resulted in salaries offered by the Ministry of Defence being around 50 per cent of those offered by commercial enterprises. This situation cannot be allowed to continue. (Paragraph 237)

31. *The Government must solve this problem. It must be able to deploy sufficient qualified staff to work on AI and to deliver demanding scrutiny of procurement offers from private developers and manufacturers. This might be achieved by establishing new pay scales, or by bringing in private sector staff on secondment. Either way, it will be challenging but absolutely necessary if we are to have the ability to compete on the international stage and safeguard our country.* (Paragraph 238)

32. *The Government has asserted that transparency and challenge are central to its approach. From the evidence we have taken, we have not found this yet to be the case. The Government should increase the transparency of advice provided by the AI Ethics Advisory Panel by publishing its Terms of Reference, membership, agendas, and minutes, as well as an annual transparency report.* (Paragraph 247)

33. *The Government should immediately expand the remit of the AI Ethics Advisory Panel to review the practical application of ethical principles in armed conflict and to cover ethics in relation to the development and use of AI in AWS.* (Paragraph 248)

34. *The Government must allow sufficient space in the Parliamentary timetable and provide enough information for Parliament, including its select committees, to scrutinise its policy on AI effectively. We naturally understand that elements of policy development may be highly sensitive; but there are established ways of dealing with such information. Arguments of secrecy must not be used to sidestep accountability.* (Paragraph 250)

35. *The Government must ensure that it engages with the public on AI-enabled AWS. It must also ensure that ethics are at the heart of its policy.* (Paragraph 251)

36. Overall, we welcome the fact that the Government has recognised the role of responsible AI in its future defence capability. AI has the potential to provide key battlefield and strategic benefits. However, in doing so, the Government must embed ethical and legal principles at all stages of design, development and deployment. Technology should be used when advantageous, but not at unacceptable cost to the UK's moral principles. (Paragraph 252)

APPENDIX 1: LIST OF MEMBERS AND DECLARATIONS OF INTEREST

List of Members

Lord Lisvane
Baroness Anderson of Stoke-on-Trent (to April 2023)
Lord Browne of Ladyton
Viscount Camrose (to March 2023)
Lord Clement-Jones
Lord Bishop of Coventry (to November 2023)
Baroness Doocey
Lord Fairfax of Cameron
Lord Grocott
Lord Hamilton of Epsom
Baroness Hodgson of Abinger
Lord Houghton of Richmond
Lord Mitchell (from April 2023)
Lord Sarfraz (from April 2023)
Baroness Symons of Vernham Dean (to April 2023)
Lord Triesman (from April 2023)

Declarations of interest

Lord Lisvane
No relevant interest
Baroness Anderson of Stoke-on-Trent (to April 2023)
Hon. Captain, Royal Navy
Member of the Advisory Board, Royal Navy Strategic Studies Centre
Board Member, Staffordshire University
Lord Browne of Ladyton
Consultant, Nuclear Threat Initiative
Director and Trustee, European Leadership Network
Board Member, Asia-pacific Network
Member of the Group of Eminent persons (GEM) CTBTO
Board Member and Vice-Chair, Nuclear Threat Initiative
Member, the Top Level Group for Multilateral Nuclear Disarmament and Non-proliferation
Ambassador, HALO Trust
Viscount Camrose (to March 2023)
Collaborated with Palantir on project work during 2022
Shareholder, Reaction Engines Limited
Lord Clement-Jones
Adviser on AI policy and regulation, DLA Piper UK LLP (law firm)
Lord Bishop of Coventry (to November 2023)
No relevant interest
Baroness Doocey
No relevant interest
Lord Fairfax of Cameron
Co-owner Hawk-i Worldwide Ltd (Private Security Company)
Lord Grocott
No relevant interest

Lord Hamilton of Epsom

Herald Investment Trust (technology companies)

As Vice Patron of the Defence Forum the Member attends dinners at the Houses of Parliament sponsored by the Defence Forum and paid for by defence industry sponsors

Baroness Hodgson of Abinger

As a member of the All-Parliamentary Parliamentary Group for the Armed Forces, the Member receives invitations to breakfast and dinner briefings which during the course of the calendar year together amount to more than £300 in value and all of which are paid for by sources listed in the register of APPGs

Trustee, Armed Forces Parliamentary Trust

Honorary Colonel, Outreach Group 11 Security Force Assistance Brigade

Lord Houghton of Richmond

Chairman, Byzgen (tech start-up)

Chairman, Draken (Europe) and FR Aviation (aviation services company)

Chairman, SecureCloud+ (technology company specialising in secure communication and data exploitation services)

Senior Defence and Security Adviser, Thales UK (defence and technical company)

Strategic Adviser, Tadaweb (technology company specialising in open source data search) (involves indirect payment for work contracted to various agencies of French and Canadian Governments)

Strategic Adviser, Whitespace (tech company specialising in AI assisted decision making)

Strategic Advisor, Rebellion Defense (tech security company)

Advisor, Draken International LLC (aviation services company) (involves indirect payment for range of contracted work including United States Air Force, Royal Saudi Air Force and Dutch Caribbean)

The Member participates in an annual dialogue on security strategy under the auspices of the Israeli Institute for National Security Studies

Strategic Adviser, Kearney (advising international governments on security and defence reform) (involves indirect payment for contracted work advising Ministry of National Guard of Saudi Arabia) (interest ceased 21 September 2022)

Lord Mitchell (from April 2023)

No relevant interest

Lord Sarfraz (from April 2023)

Venture Partner, Draper Associates (early-stage technology venture capital)

Member of the Advisory Board, C3 AI (enterprise artificial intelligence software)

Baroness Symons of Vernham Dean (to April 2023)

No relevant interest

Lord Triesman (from April 2023)

No relevant interest

A full list of Members’ interests can be found in the Register of Lords’ interests: https://members.parliament.uk/members/lords/interests/register-of-lords-interests

Specialist Advisers

Dr Adrian Weller

Director of Research, Machine Learning, University of Cambridge

Head of Safe and Ethical AI, The Alan Turing Institute

Adviser to retrain.ai, an artificial intelligence recruitment platform, for which he receives shares/options

Professor Dame Muffy Calder (from September 2023)

Vice Principal and Head of College of Science and Engineering, University of Glasgow

Chair, Technology Advisory Panel, Investigatory Powers Commissioner's Office

Member, Prime Minister's Council for Science and Technology

APPENDIX 2: LIST OF WITNESSES

Evidence is published online at https://committees.parliament.uk/committee/646/ai-in-weapon-systems-committee/publications/ and is available for inspection at the Parliamentary Archives (020 7219 3074).

Evidence received by the Committee is listed below in chronological order of oral evidence session, and then in alphabetical order. Those witnesses marked with ** gave both oral evidence and written evidence. Those marked with * gave oral evidence and did not submit any written evidence. All other witnesses submitted written evidence only.

Oral evidence in chronological order

*	Professor Noam Lubell, Professor, University of Essex School of Law	QQ 1–14
*	Georgia Hinds, Legal Adviser, International Committee of the Red Cross (ICRC)	
*	Dr Daragh Murray, Senior Lecturer and IHSS Fellow, Queen Mary University of London School of Law	
*	Yasmin Afina, Research Associate, Chatham House	QQ 15–23
*	Dr 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	
**	Courtney Bowman, Global Director of Privacy and Civil Liberties Engineering, Palantir Technologies UK	QQ 24–42
*	Professor Kenneth Payne, Professor of Strategy, Kings College London	
*	James Black, Assistant Director of the Defence and Security Research Group, RAND Group	
*	Dr Keith Dear, Director of Artificial Intelligence Innovation, Fujitsu	
*	Professor Mariarosaria Taddeo, Associate Professor, Oxford Internet Institute	QQ 43–63
*	Dr Alexander Blanchard, Digital Ethics Research Fellow, Alan Turing Institute	
*	Verity Coyle, Senior Advisor, Amnesty International	
*	Professor Sir Lawrence Freedman, Emeritus Professor of War Studies, King's College London	QQ 64–80

★	Dr Jurriaan van Diggelen, Senior Researcher in AI and Program Leader, Human-machine Teaming, Netherlands Organisation for Applied Scientific Research	QQ 81–95
★	Professor Dame Muffy Calder, Vice Principal and Head of College of Science and Engineering, University of Glasgow	
★	Professor Gopal Ramchurn, Professor of Artificial Intelligence, University of Southampton	
★	Lord Sedwill	QQ 96–108
★	Professor Hugh Durrant-Whyte, Director of the Centre for Translation Data Science, University of Sydney	
★★	Richard Moyes, Managing Director, Article 36	QQ 109–119
★	Professor Noel Sharkey, Emeritus Professor of AI and Robotics and Professor of Public Engagement, University of Sheffield	
★	Dr Paddy Walker, Senior Research Fellow in Modern War Studies, University of Buckingham	
★	Professor Stuart Russell, Professor of Computer Science, University of California, Berkeley	QQ 120–132
★	Dr James Johnson, Lecturer in Strategic Studies, University of Aberdeen	QQ 133–138
★	Christopher King, Head of Weapons of Mass Destruction Branch, UN Office for Disarmament Affairs	
★	General Sir Chris Deverell, Former Commander, Joint Forces Command	QQ 139–155
★★	Tsvetelina Van Benthem, DPhil, Candidate in Public International Law, University of Oxford	
★	Laura Nolan, SRE and Principal Engineer, Stanza Systems	QQ 156–164
★★	Taniel Yusef, Visiting Researcher, Centre for the Study of Existential Risk	
★★	Professor Christian Enemark, Professor of International Relations, University of Southampton	
★★	James Cartlidge MP, Minister for Defence Procurement, Ministry of Defence	QQ 165–190
★★	Paul Lincoln, Second Permanent Secretary, Ministry of Defence	
★★	Lieutenant General Tom Copinger-Symes, Deputy Commander UK Strategic Command, Ministry of Defence	

*	Professor Jinghan Zeng, Professor of China and International Studies, Lancaster University	QQ 191–200
*	Andrew Kinniburgh, Director General, Make UK Defence	QQ 201–207

Alphabetical list of witnesses

	Dr David Anderson	AIW0041
	Stuart Anderson	AIW0042
	Anduril Industries	AIW0011
	Anonymous	AIW0030
	ART–AI	AIW0016
**	Article 36 (QQ 109–119)	AIW0017
	BAE Systems	AIW0022
*	Dr Alexander Blanchard, Digital Ethics Research Fellow, Alan Turing Institute (QQ 43–63)	
	Dr Ingvild Bode	AIW0015
	Professor William Boothby	AIW0003
	Dr Emma Breeze	AIW0007
	British Institute of Technology Ltd	AIW0028
*	Professor Dame Muffy Calder (QQ 81–95)	
*	Chatham House (QQ 15–23)	
	Professor Thompson Chengeta	AIW0020
	Professor Thompson Chengeta	AIW0021
*	Verity Coyle, Senior Advisor, Amnesty International (QQ 43–63)	
*	Dr Keith Dear (QQ 24–42)	
	DeepMind	AIW0037
*	General Sir Chris Deverell (QQ 139–155)	
	Drone Wars UK	AIW0008
*	Professor Hugh Durrant-Whyte (QQ 96–108)	
**	Professor Christian Enemark (QQ 156–164)	AIW0004
	European Leadership Network	AIW0023
	Dr Mikolaj Firlej	AIW0034
*	Sir Lawrence Freedman (QQ 64–80)	
	GCH Technologies	AIW0024
	Professor Steven Haines	AIW0032
	Rebecca Hall	AIW0013

	Dr Stephen Harwood	AIW0010
	Francis Heritage	AIW0029
★	International Committee of the Red Cross (ICRC) (QQ 1–14)	
★	Dr James Johnson (QQ 133–138)	
	Nicolas Jouan	AIW0040
	Maximilian Kiener	AIW0012
★	Christopher King (QQ 133–138)	
	Labour for the Long Term	AIW0031
★	Professor Noam Lubell (QQ 1–14)	
★	Make UK Defence (QQ 201–207)	
★★	Ministry of Defence (QQ 165–190)	AIW0035
★	Dr Daragh Murray (QQ 1–14)	
★	Laura Nolan (QQ 156–164)	
	Andrew Otterbacher	AIW0043
★	Charles Ovink (QQ 15–23)	
★★	Palantir UK (QQ 24–42)	AIW0025
★	Dr Kenneth Payne (QQ 24–42)	
	Peace Research Institute Frankfurt	AIW0002
★	Professor Gopal Ramchurn (QQ 81–95)	
★	RAND Europe (QQ 24–42)	
★	Professor Stuart Russell (QQ 24–42)	
	Dr Elke Schwarz	AIW0009
★	Lord Sedwill (QQ 97–108)	
★	Professor Noel Sharkey (QQ 109–119)	
	Peter Spayne	AIW0005
★	Stockholm International Peace Research Institute (QQ 15–23)	
	Stop Killer Robots	AIW0018
★	Professor Mariarosaria Taddeo (QQ 43–63)	
	UK Campaign to Stop Killer Robots	AIW0038
	Dr Ozlem Ulgen	AIW0019
★★	Tsvetalina Van Benthem (QQ 139–155)	AIW0033
★	Dr Jurriaan Van Diggelen (QQ 81–95)	
★	Paddy Walker (QQ 109–119)	
	Dr Rodrick Wallace	AIW0027
	Professor Toby Walsh	AIW0026

	Dr Tom Watts	AIW0014
	Elliot Winter	AIW0001
	Women's International League for Peace and Freedom	AIW0006
**	Taniel Yusef (QQ 156–164)	AIW0039
*	Professor Jinghan Zeng (QQ 191–200)	

APPENDIX 3: CALL FOR EVIDENCE

Aim of the inquiry

The Artificial Intelligence in Weapon Systems Committee was appointed on 31 January 2023 to consider the use of artificial intelligence in weapon systems.

Automation refers to the use of systems to perform tasks that would ordinarily involve human input. Automation and autonomy can be viewed as existing on a spectrum relating to the level of human supervision over a system. This can range from manually controlled systems to those that independently make decisions about how to achieve certain human-set goals. Autonomy is a characteristic of a system using artificial intelligence to determine its own course of action by making its own decisions.

Autonomous weapons systems (AWS), also known as lethal autonomous weapons systems (LAWS), are weapons systems which can select, detect and engage targets with little to no human intervention. The scope of these systems can vary significantly, from fully autonomous weapons that can operate without any human involvement, to semi-autonomous weapons that require human action to launch an attack. The UK does not currently have an operative definition of AWS.

The House has asked the Committee to conclude its inquiry by the end of November 2023. The Government has undertaken to respond in writing to select committee reports, usually within two months of publication.

The Committee expects to hear from invited contributors in public sessions from March to July 2023 inclusive.

This is a public call for written evidence to be submitted to the Committee. The deadline is Monday 10 April 2023.

The Committee is happy to receive submissions on any issues related to artificial intelligence in weapons systems but would particularly welcome submissions on the questions listed below.

Contributors need not address every question and experts are encouraged to focus on their specialism. Other issues may be discussed provided that their relevance is explained. Submissions which have been previously published will not be accepted as evidence. However, published material may be referenced where relevant.

The Committee encourages people from all backgrounds to contribute and believes that it is particularly important to hear from groups which are often under-represented. The Committee's work is most effective when it is informed by as diverse a range of perspectives and experiences as possible. Please pass this on to others who may be interested in contributing.

Instructions on how to submit evidence are set out at the end of this document. If you have any queries please email the staff of the Committee at hlaiweapons@parliament.uk. When preparing your response, please bear in mind that short, concise submissions are preferred. Please explain any acronyms or technical terms, and ensure your submission is understandable by a lay audience.

Questions

1. What do you understand by the term autonomous weapons system (AWS)? Should the UK adopt an operative definition of AWS?

2. What are the possible challenges, risks and benefits of AWS? How would AWS change the makeup of defence forces and the nature of combat?

3. What safeguards (technological, legal or otherwise) would be needed to ensure safe, reliable and accountable AWS?

4. Is existing international humanitarian law (IHL) sufficient to ensure any AWS act safely and appropriately? What oversight or accountability measures are necessary to ensure compliance with IHL? If IHL is insufficient, what other mechanisms should be introduced to regulate AWS?

5. What are your views on the Government's AI Defence Strategy[418] and the policy statement 'Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence'?[419] Are these sufficient in guiding the development and application of AWS?

[418] MoD, 'Defence Artificial Intelligence Strategy' (15 June 2022): https://www.gov.uk/government/publications/defence-artificial-intelligence-strategy/defence-artificial-intelligence-strategy [accessed 17 February 2023]

[419] MoD, *Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence* (15 June 2022): https://www.gov.uk/government/publications/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence/ambitious-safe-responsible-our-approach-to-the-delivery-of-ai-enabled-capability-in-defence [accessed 17 February 2023]

APPENDIX 4: EXAMPLES OF AWS AND AI-ENABLED WEAPONS

AI in development or testing phases

Startle and Sycoiea (UK): These two AI systems work together and have been developed for the Royal Navy. Startle is an autonomous threat monitoring system that assists sailors in the operations room. It provides live recommendations and alerts and aims to enable more rapid reaction to threats. Sycoiea builds on Startle by identifying the nearest threat and recommending the best weapon to deal with it. Both were tested on HMS Lancaster and HMS Dragon during the NATO At Sea Demonstration/Formidable Shield 2021 (ASD/FS-21) live fire exercise as part of air and missile defence.[420]

Naval Strike Missile (USA): Developed by Raytheon, the missile uses an AI-based guidance system to improve precision for over-the-horizon defence. It has a range of over 100 nautical miles and seems to be fully autonomous once deployed, with no more human interference.[421]

Robotic Combat Vehicle-Light (USA): Autonomous light armoured vehicles developed for teaming with manned vehicles. Tested at Ford Hood in 2022. Vehicles can operate autonomously, semi-autonomously, or be remotely controlled. The Army hopes that AI technology will allow a single operator to control multiple RCVs. The light vehicle prototypes are equipped with a 25mm main gun on a remote turret and further weapons are being tested including the M153 Common Remotely Operated Weapons Station II (CROWS II), the 0.50-calibre M2 machine gun, and the 40mm MK19 Mod 3 automatic grenade launcher. It is not clear what systems on the vehicle involve AI.[422]

Boing MQ-28A Ghost Bat (Australia): Also known as 'Loyal Wingman', the Ghost Bat flies alongside manned aircraft and uses AI to perform crewed-uncrewed teaming missions. It currently performs surveillance, reconnaissance, and early warning missions. It is planned to enter into RAAF service in 2024–25.[423]

Next Generation Air Dominance (USA): System-of-systems approach that integrates air power assets with manned 5th and 6th generation fighter jets accompanied by wingman-style UAVs, being called collaborative combat aircraft (CCAs), and potentially other assets. AI is an integral component of the system to enable teaming between the manned fighters and CCAs and to provide comprehensive situational awareness, improved survivability, and greater lethality. CCAs may be used as a shooters, jammers, or sensors. CCAs may also be used as an entirely autonomous platform, like a swarm, potentially without direct

[420] MoD, 'Artificial Intelligence used at sea for first time', (29 May 2021): https://www.gov.uk/government/news/artificial-intelligence-used-at-sea-for-first-time [accessed 28 July 2023]. Royal Navy, 'Navy tests artificial intelligence against supersonic missiles' (29 May 2021): https://www.royalnavy.mod.uk/news-and-latest-activity/news/2021/may/29/20210529-artificial-intelligence [accessed 28 July 2023]. Roke, 'Roke to showcase STARTLE AI capability as part of Formidable Shield 2021' (18 May 2021): https://roke.co.uk/news/roke-to-showcase-startle-ai-capability-at-formidable-shield [accessed 28 July 2023]

[421] Emerj, 'Artificial Intelligence in the Navy – Current Contractors and Innovations' (23 January 2019): https://emerj.com/ai-sector-overviews/ai-in-the-navy-current-contractors-and-innovations/ [accessed 28 July 2023]

[422] Congressional Research Service, 'The Army's Robotic Combat Vehicle (RCV) Program' (3 April 2023): https://crsreports.congress.gov/product/pdf/IF/IF11876 [accessed 28 July 2023] and Army Technology, 'Robotic Combat Vehicle-Light (RCV-L) (22 May 2021): https://www.army-technology.com/projects/robotic-combat-vehicle-light-rcv-l/ [accessed 28 July 2023]

[423] Airforce Technology, 'MQ-28A Ghost Bat Unmanned Aircraft, Australia' (22 June 2023): https://www.airforce-technology.com/projects/loyal-wingman-unmanned-aircraft/ [accessed 28 July 2023]

human supervision and for missions which could include striking targets. The full capability platform is not expected to enter service until 2030.[424]

Air Combat Evolution (USA): Being developed by DARPA, ACE aims to enable the pilot of a manned aircraft to command a more global mission, rather than operating a single platform, through teaming with unmanned systems that can engage in individual tactics. They have set human-machine collaboration in aircraft dog-fighting as their core challenge. The programme has completed virtual simulations pitting AI against AI and AI against an experienced F-16 fighter pilot. The system has also successfully completed tests using the VISTA X-62A test aircraft out of Edwards Air Force Base. In the test, the system was able to execute within-visual-range tactical manoeuvring against AI red-team agents.[425]

Autonomous Air Combat Operations (USA): Within the Air Force Research Laboratory, AACO is developing an AI piloting system capable of advanced intelligence, surveillance, and reconnaissance (ISR) as well as beyond-visual-range combat. They have also carried out tests with the VISTA X-62A out of Edwards Air Force Base, which focused on combat with a single adversary beyond visual range.[426]

Skyborg (USA): Another effort to develop an unmanned 'loyal wingman' for human pilots. The aim is for the system to be integrated into a range of drones that can engage in missions too risky for human pilots.[427]

Elbit Swarming Drones (UK): The British Army successfully trialled this capability in September 2022 through its nano-unmanned aerial systems project. One of the demonstrations was by Elbit Systems, an Israeli company, where they used AI so that one operator could control six drones at once. The operator set an overarching surveillance goal and the system then autonomously set the missions/tasks of each drone. This has not yet been weaponised.[428]

[424] Airforce Technology, 'Next Generation Air Dominance Programme' (13 July 2023): https://www.airforce-technology.com/projects/next-generation-air-dominance-programme-us/ [accessed 28 July 2023]. Airforce Technology, 'Collaborative Combat Aircraft (CCA), USA' (17 July 2023): https://www.airforce-technology.com/projects/collaborative-combat-aircraft-cca-usa/ [accessed 28 July 2023]. National Defense, 'Air Force Putting Software First for Next-Gen Air Dominance (Updated)' (29 July 2022): https://www.nationaldefensemagazine.org/articles/2022/7/29/air-force-putting-software-first-for-next-gen-air-dominance [accessed 28 July 2023] and Airforce Technology, 'GA-ASI demonstrates multiple UAS missions with AI pilots' (12 January 2023): https://www.airforce-technology.com/news/gaasi-uas-ai-pilots/ [accessed 28 July 2023]

[425] DARPA, 'Air Combat Evolution (ACE)': https://www.darpa.mil/program/air-combat-evolution [accessed 28 July 2023] and The Aviationist, 'Artificial Intelligence Successfully Piloted The X-62 VISTA' (14 February 2023): https://theaviationist.com/2023/02/14/artificial-intelligence-successfully-piloted-the-x-62-vista/ [accessed 28 July 2023]

[426] Wired, 'The US Air Force Is Moving Fast on AI-Piloted Fighter Jets' (8 March 2023): https://www.wired.com/story/us-air-force-skyborg-vista-ai-fighter-jets/ [accessed 28 July 2023]

[427] AFRL, 'Syborg': https://afresearchlab.com/technology/vanguards/successstories/skyborg [accessed 28 July 2023] and Defense News, 'The Air Force's first Skyborg autonomous drone prototype made its first flight' (5 May 2021): https://www.defensenews.com/air/2021/05/05/the-air-forces-first-skyborg-autonomous-drone-prototype-made-its-first-flight/ [accessed 28 July 2023]

[428] Aerospace Manufacturing, 'Elbit Systems UK to deliver drone swarms to MoD' (8 April 2022): https://www.aero-mag.com/elbit-systems-drone-swarm-08042022 [accessed 28 July 2023] and British Army, 'British Army carries out successful Swarming Drone capability' (8 September 2022): https://www.army.mod.uk/news-and-events/news/2022/09/british-army-carries-out-successful-swarming-drone-capability/ [accessed 28 July 2023].

Examples of existing UK systems with degrees of autonomy

Brimstone Missile: Can be programmed to search, identify, track, and strike vehicles using sensor data. It is produced by MBDA Missile Systems, which is an integrated subsidiary of Airbus, BAE Systems, and Leonardo.[429]

MQ-9 Reaper: The only armed drone in the UK arsenal that is capable of autonomous flight. It is produced by General Atomics Aeronautical Systems.[430]

Phalanx Close-in Weapon System: This is probably the oldest autonomous weapon system. It defends military ships from incoming threats via its own target identification system. It has been in use since 1973. The Royal Navy uses it widely, including on the aircraft carrier HMS Queen Elizabeth.[431]

Examples of systems with degrees of autonomy used by other states

HARPY loitering munition: This is an all-weather day/night 'Fire and Forget' autonomous weapon, launched from a ground vehicle behind the battle zone. They are programmed before launch to perform autonomous flight to a pre-defined "Loitering Area", in which they loiter and search for radiating targets.[432] There are no apparent instances of use in combat, although a larger version - the Harop - has been used by Azerbaijan, Israel, and Morocco.[433]

KARGU: This is a portable, rotary wing attack drone designed to provide tactical ISR and precision strike capabilities for ground troops. KARGU can navigate autonomously, but requires human intervention to target.[434] Used by Turkey.[435]

[429] Parliamentary Office of Science and Technology, *Automation in Military Operations*, No 681 (October 2022): https://researchbriefings.files.parliament.uk/documents/POST-PN-0681/POST-PN-0681.pdf

[430] *Ibid.*

[431] *Ibid.*

[432] Israel Aerospace Industries, 'HARPY Autonomous Weapon for All Weather': https://www.iai.co.il/p/harpy [accessed 20 October 2023]

[433] Military Today, 'IAI Harop': https://www.militarytoday.com/aircraft/harop.htm [accessed 20 October 2023]

[434] STM, 'KARGU: Combat Proven Rotary Wing Loitering Munition System': https://www.stm.com.tr/en/kargu-autonomous-tactical-multi-rotor-attack-uav [accessed 20 October 2023]

[435] Forbes, 'Turkish Military To Receive 500 Swarming Kamikaze Drones' (17 June 2020): https://www.forbes.com/sites/davidhambling/2020/06/17/turkish-military-to-receive-500-swarming-kamikaze-drones/?sh=2a3a8d56251a [accessed 20 October 2023]

APPENDIX 5: GLOSSARY

Article 36 reviews

Contained in Article 36 of Additional Protocol I of the Geneva convention, this requires states party to the Convention "in the study, development, acquisition or adoption of a new weapon, means or method of warfare … to determine whether its employment would, in some or all circumstances, be prohibited by [Additional Protocol I or other applicable international law]."[436]

Artificial intelligence (AI)

"a family of general-purpose technologies, any of which may enable machines to perform tasks normally requiring human or biological intelligence, especially when the machines learn from data how to do those tasks."[437] AI can be either 'general' or 'narrow'. Artificial general intelligence refers to a machine with broad cognitive abilities, which is able to think, or at least simulate convincingly, many or all of the intellectual capacities of a human being, and potentially surpass them. By contrast, narrow AI systems perform specific tasks in a limited domain, which would require intelligence in a being and may even surpass human abilities in these areas. However, such systems are limited in the range of tasks they can perform.

Automation

The use of systems to perform tasks that would ordinarily involve human input.

Autonomous weapon system (AWS)

Weapon systems which can detect, select and engage targets with little to no human intervention or possess some degree of autonomy in one or more aspect. See glossary entries for 'fully autonomous weapon systems' and 'partially autonomous weapon systems'.

Benchmarking

Standardised tests that measure the performance of systems on specific tasks.

Black box AI

An AI system that operates in a way which is not readily visible or intelligible

Distinction (principle of international humanitarian law)

Parties to a conflict must at all times distinguish between civilians and combatants, and between civilian objects and military objectives. Attacks must be directed solely at combatants or military objectives and attacks that fail to distinguish between civilians and combatants are classified as indiscriminate and unlawful.

[436] International Committee of the Red Cross, 'Protocols Additional to the Geneva Conventions of 12 August 1949', p 30: https://www.icrc.org/en/doc/assets/files/other/icrc_002_0321.pdf [accessed 3 October 2023]

[437] MoD, *Defence Artificial Intelligence Strategy* (June 2022): https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1082416/Defence_Artificial_Intelligence_Strategy.pdf [accessed 3 October 2023]

Explainable AI	AI that delivers accompanying relevant evidence or reasons for outcomes and processes.[438]
Fully autonomous weapon systems	Systems that, once activated, can identify, select, and engage targets with lethal force without further intervention by an operator.
Humanity (principle of international humanitarian law)	The principle of humanity forbids a party to a conflict from imposing any suffering, injury or destruction which is not necessary to achieve legitimate military purposes.
International humanitarian law (IHL)	A set of rules which seek, for humanitarian reasons, to limit the effects of armed conflict.[439]
Machine learning (ML)	A subfield of AI that uses data, algorithms and statistical models to learn patterns, derive insights, and make predictions.
Military necessity (principle of international humanitarian law)	Military necessity dictates that military force should only be used against the enemy to the extent necessary to achieve a legitimate purpose of the conflict.
Neural networks	Layers of artificial neurons used to learn patterns and representations of data and output predictions or decisions based on what they have learned.
Open-source software	Software released under a licence in which the copyright holder grants users the right to use, change and distribute the source code to any person for any reason.
Partially autonomous weapon systems	Systems featuring varying degrees of decision-making autonomy in critical functions such as identification, classification, interception and engagement.
Predictability of an AI system	The ability (of a human) to predict or reason about the outputs of the system, with varying degrees of certainty.

[438] Phillips et al., US Department of Commerce, National Institute of Standards and Technology, *Four Principles of Explainable Artificial Intelligence*, (September 2021): https://doi.org/10.6028/NIST.IR.8312 [accessed 3 October 2023]

[439] International Committee of the Red Cross, *What is International Humanitarian Law* (July 2004): https://www.icrc.org/en/doc/assets/files/other/what_is_ihl.pdf [accessed 24 November 2023]

Proportionality (principle of international humanitarian law)

IHL does not prohibit attacks which may cause incidental harm to civilians or civilian objects, but attacks which cause disproportionate civilian harm relative to the military benefits are unlawful. Additional Protocol I to the Geneva Conventions defines a disproportionate attack as one that "may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof, which would be excessive in relation to the concrete and direct military advantage anticipated."

Reliability

The ability of an AI system to perform a required function reliably under stated conditions for a stated time.

Synthetic data

Data that is made up, by humans or generated by machine, and may be representative of real-world data or meet certain conditions. One use of synthetic data is to stress test systems for difficult scenarios that might occur. As with real data, this may not always be accurate.

Training data

Data used during the process of training a machine learning algorithm.

Transparency

The ability for a user or interested party to understand important aspects of an AI system, often including how it makes decisions and processes data.

UN Convention on Certain Conventional Weapons

A UN Convention with the purpose to ban or restrict the use of specific types of weapons that are considered to cause unnecessary or unjustifiable suffering to combatants or to affect civilians indiscriminately, also known as the 'Inhumane Weapons Convention'.[440]

[440] UN Office for Disarmament Affairs, 'The Convention on Certain Conventional Weapons': https://disarmament.unoda.org/the-convention-on-certain-conventional-weapons/ [accessed 26 September 2023]