Professor Stuart Anderson, Professor of Dependable Systems, The University of Edinburgh, Professor Alex Lascarides, Chair in Semantics, The University of Edinburgh, Professor Subramanian (Ram) Ramamoorthy, Personal Chair of Robot Learning and Autonomy, The University of Edinburgh, Professor Shannon Vallor, Baillie Gifford Chair of the Ethics of Data and Artificial Intelligence, The University of Edinburgh - Written Evidence (AIW0042)

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

    1.1. The UKRI Trustworthy Autonomous Systems programme, on which several of us work, defines an autonomous system as "A system involving software applications, machines, and people, that is able to take actions with little or no human supervision." (https://tas.ac.uk/our-definitions) Autonomous weapons systems are often divided into those that can initiate kinetic actions of a potentially **lethal** nature with little or no human supervision (LAWS) and the broader category of AWS. The latter may include nonlethal autonomous weapons (such as taser-equipped drones) as well as autonomous components of a weapons system that are precursors to lethal action (for example, autonomous targeting in a weapons system for which firing decisions remain under sole human control). Autonomy in AWS is a matter of degree; that is, AWS exist along a spectrum of machine autonomy, inversely proportional to the amount of human instruction or supervision they require in order to carry out their designed objective(s).

    1.2. There is, however, another meaning of 'autonomous' in the AWS context, distinct from autonomy in the sense defined in 1.1. The second, moral and philosophical meaning of 'autonomy' refers to the capacity of a thinking agent to choose their actions with intent, in accordance with understanding and deliberative reason, rather than acting under physical compulsion. Both meanings of ''autonomy' appear frequently in the academic and policy literature on the ethical and legal implications of AWS, particularly LAWS. This is because autonomy in this second sense is often argued to be a requirement for moral and/or legal responsibility, including command responsibility. Many

scholars assert that human rights and international law require any actions taken by a LAWS to be attributable to a responsible and accountable human party or parties, who have the rational decision-making capacity of an autonomous moral agent (see Marchant et al 2011).

1.3. It is therefore helpful for clarity's sake to speak either of 'system autonomy' or ''moral autonomy,' to indicate which notion we mean in a given AWS context. There is a strong consensus among AI researchers and legal scholars alike that today's capacities for *system autonomy* (also called 'machine autonomy'), do not include or enable the *moral autonomy* required to assume moral or legal responsibility for AWS actions and decisions. This holds true even for the most advanced AWS in development, as they merely execute mathematical *procedures* rather than weigh moral and legal *reasons*. There are, then, no autonomous systems in the second sense of autonomy defined in 1.2, nor is there any evidence of such systems on the horizon. Speculative claims that LLMs (large language models) exhibit capacities of 'AGI' (artificial general intelligence) such as conscious intention, reasoning and understanding are false. LLMs use computational techniques to select statistically plausible chains of language, without the need or ability to grasp the meaning expressed in that language. LLMs are incapable of understanding what the concepts of justice, military necessity, proportionality, human dignity or death signify in the human experience of the world, as these systems have no 'world' to experience, only mathematical matrices.

1.4. In short, then, *system autonomy* in AWS as a real capability does not alleviate the need for human accountability, since AWS for the foreseeable future will still lack the *moral autonomy* to be judged responsible for the actions they initiate or the 'decisions' (more properly, computations) they perform.

1.5. Yet the ability of AWS to initiate action without human supervision creates an appearance of so-*called responsibility gaps* that demand legal and policy solutions. Responsibility gaps arise most urgently in AWS enabled by AI technologies that can 'learn' new behaviours or devise new strategies in real time, in a real-world environment. If an AWS with such capabilities takes an action as a result of a set of calculations that no human designer or deployer predicted, due to environmental influences and random

(stochastic) learning strategies, and subsequent post-hoc analysis fails to establish that the behaviour was predictable, it may be difficult to assign 'fault' for an undesirable or unjust outcome.

1.6. In response to the responsibility gap challenge, Vallor and Ganesh (see reference 1 below) have developed a *Responsibility Framework* calling for regulatory and other governance interventions to ensure that the powers of any high-stakes autonomous system (as defined in 1.1) are always balanced by strict and specific human duties of care. These must be assigned to identifiable and responsible human agents who are prepared to be liable or otherwise answerable for the consequences of unexpected and undesirable behaviours of a system. This may perhaps provide the framework to link these considerations of IHL with the technical requirements of defence organisations noted below.

1.7. In answer to the second question, it would be helpful for any UK operational definition to make clear that AWS are autonomous only in the limited mechanical sense of an agent that can act without human direction or supervision (1.1), as distinct from the rational/moral/legal autonomy (1.2) that AWS continue to lack.

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

2.1. **Risks and Ethical Concerns:**

AWS present numerous ethical and legal concerns, including:

2.1.1 **potential violations of international human rights and humanitarian law**, as well as **violations of the Laws of Armed Conflict emanating from just war theory** that mandate that warfighters observe the principles of proportionality, necessity and distinction. While there remains debate about whether one has a human right not to be killed by an autonomous machine, many legal scholars and philosophers have asserted that the application of just war principles presupposes the rational moral agency that AWS lack, and that combatants killed by an AWS which cannot weigh these principles is a prima facie moral and legal wrong.

2.1.2 **unmanageable failures of human control of AWS,** which could lead to unacceptable deaths of noncombatants or catastrophic 'friendly fire' incidents, either by malfunction, unpredictable learning strategies (also known as 'reward

hacking'), or malicious infiltration of AWS by hackers or other bad actors, for example via viruses or adversarial inputs designed to disrupt the system's intended function.

2.1.3 **fostering of a 'race to the bottom' toward ever more dangerous and uncontrollable AWS capabilities,** for example if nation-states see AWS as an opportunity to gain a strategic advantage and refuse to delay deployment until proper testing, safeguards and legal guardrails are established.

2.1.4 **incentivizing ever greater departures from modern norms of warfighting by pursuing AWS as a means of asymmetrical warfare;** the concern here is that war may become more dangerous for all if nations unable to compete in AWS capabilities perceive this new mode of military action to be so unfair and 'inhuman' as to justify dispensing with any restraint; for example, in pursuing biological or chemical weapons or attacks on civilians in order to 'level the playing field.' See P.W. Singer's (2009) *Wired for War*, as well as Skerker et al (2020), who explore the idea of AWS as violating an implicit 'martial contract' between combatants, risking unanticipated harms and norm degradation: https://link.springer.com/article/10.1007/s10676-020-09528-0

2.1.5. **potential bias and discrimination in AWS** as we have seen occur in virtually all other domains of automated AI decision-making, from benefits fraud detection algorithms to healthcare resource allocation algorithms to predictive policing algorithms. Each of these domains of AI application invariably involve datasets infected with human racial, gender, class and ethnic biases, signals which the models tend to pick up and amplify in their outputs, creating a 'runaway feedback loop' where human social bias feeds AI bias, which in turn perpetuates and amplifies the original social bias. If AWS models are trained on data from biased human warfighters (for example, those whose judgments of 'likely combatant' or 'not likely combatant' are influenced by ethnic or religious stereotypes), or if AWS data is labeled by human workers with similar biases, the models will inherit and amplify those unjust biases in their outputs.

2.1.6. **potential moral deskilling of the military profession**, as norms of military virtue and excellence may be weakened by the greater delegation of high-stakes decision making to AWS (see

https://www.ccdcoe.org/uploads/2018/10/9_d2r1s10_vallor.pdf) A parallel risk, though not specifically a moral one, is the strategic deskilling of military judgment as human opportunities to develop, implement and learn from strategic decisions are lost.

2.2. **Potential Benefits:**

2.2.1 Strategic benefits of AWS are presumed to follow from their advantages of speed, scale and computational power over human decision-makers. For example, swarm algorithms can in principle coordinate and control a large squadron of UAVs and adapt to a changing battlefield situation far more quickly and efficiently than the efforts of individual human pilots to coordinate their attack in real time. Many such proposed AWS benefits remain speculative, even if intuitively plausible. Note that autonomous cars still often react very poorly to unexpected changes in the driving environment, so AWS performance is likely to be equally brittle in complex 'open-world' battle environments as compared to well-controlled simulations.

2.2.3 The earliest moral justifications claimed for AWS include the **highly speculative promise of machines that will more consistently follow the norms of warfighting** due to their lack of psychological and biological vulnerability. The idea, as laid out by Ron Arkin in 2009's *Governing Lethal Behaviour in Autonomous Robots*, has been that AWS will not be tempted to violate combatant and noncombatant rights out of anger, fear, or other mental distress, and that this makes them more trustworthy warfighters than humans. However, Arkin's argument (and those who share his view) has always been premised on the assumption that AWS can be fitted with capabilities of 'machine ethics' or an 'ethical governor' module that ensure it complies with the laws of armed conflict. Little to no progress has been made in this area of applied AI research and there is no indication on the horizon of AWS that display any reliable moral competence or even capacities for legal compliance. This promise of benefit, for the foreseeable future, *remains an entirely empty one.*

2.2.4 The most common justification for AWS is the **claim that they will enable nations to reduce net human casualties of warfighting**, since fewer humans will be needed in a given combat zone. Some argue that this moral and strategic good, which has yet to be demonstrated, can outweigh other moral considerations. This has raised questions of how the strategic purpose of military conflict

changes when humans are increasingly removed from the battle space (‘robots fighting robots’), but due to asymmetrical AWS capabilities, in reality there is little chance of that removal happening unilaterally.

2.2.5 Dealing with **very large volumes of data**. As the number and sophistication of sensors increases there is a massive increase in the volume of data available, beyond the human capacity to process. Machine learning models provide sophisticated means of summarizing very large volumes of data and they may become essential in sensemaking of large-volume, real-time data streams, which are relevant to the AWS context.

2.3 **Challenges: Complexity of Machine Learning Supply Chains:**

2.3.1 **Many Hands:** Machine learning model supply chains are long and can include many actors and companies, who are constantly shifting. Keeping track of the myriad of actors involved in a model’s development cycle appears can be virtually impossible for today’s machine learning models (see https://dl.acm.org/doi/abs/10.1145/3593013.3594073), which presents serious challenges for effective governance, safety, security and reliability of these models.

2.3.2 **Many Things** (particularly data): Today, machine learning models have uncertain provenance and training data sets have similar issues. Using operational data to train and test adds further uncertainty and additional requirements to capture, store, annotate and curate that data. There is also concern about the limited pool of training data available for the AWS context; unlike models trained on virtually the entirety of the Internet, training of AWS models will largely rely on simulations and synthetic data, which do not always deliver the results and robustness expected.

2.3.3 **Learning in Use**: One potentially important aspect of AI weapons is the proliferation of “different” systems as they learn in specific contexts of use, resulting in many different variations of the same base model. This poses governance and safety issues in keeping track of multiple versions of models, and potentially needing to deal with the safety assurance of each model on an individual basis. As noted above, ML models are also highly susceptible to dramatic performance changes with even small changes of context. In an adversarial context this is particularly problematic.

2.4 **Other Challenges:** The recent ISO/IEC 23894:2023 standard on Risk Management for AI recognized the need to reconsider the

following risk management principles. The standard sees no need to extend the principles but considers the following principles have unique features when we consider AI:

2.4.1. **Inclusive**: Machine learning systems typically have a wide range of potential stakeholders and accessing stakeholders and eliciting their view of risk is demanding.

2.4.2 **Dynamic**: As noted above, because machine learning systems can be highly context dependent, they are sensitive to context change and new risks can arise very rapidly; this means that risk assessment needs to be continuous and responsive rather than static. Factors influencing assessment of model performance may include human and cultural factors that are hard to predict and control.

2.4.3 **Best Available Information**: Responsible innovation will require that manufacturers of machine learning systems keep track of the uses and performance of their models; this potentially places restrictions on the acceptable use of such systems and may involve additional need to track changes in the system. These issues will need careful management.

2.4.4 **Continual Improvement, Interactivity and Interdependence:** Organisations operating AWS will be managing a complex, interacting, technology ecosystem where new risks may be emerging in one system that need to be considered in other systems, and there may be emergent risk from the unforseeable interactions of different machine learning systems (consider the 'flash crashes' caused earlier in the century by rapid, unexpected interactions between high frequency trading algorithms. Again, this is a significant management task.

2.5. **Additional Regulatory Challenges**: Linking the safety requirements of defence organisations to the well-established engineering approach to safety and security systems engineering may require additional regulatory effort to take account of the probabilistic aspects of machine learning components in systems. The traditional approach of safety engineering has been to require determinacy in all aspects of the system particularly for systems with very strict safety requirements (for example, the original versions of the UK 00-55 standard required formal mathematical proofs of safety). This approach is untenable for systems including sophisticated machine learning components. Within this process, it is desirable to establish an approach which integrates legal and ethical insights from the outset of the conceptualization, design, development, entry into service and decommissioning of an AI enabled weapons system. Thus, the

existing architecture of legal review, legal evaluation and legal guidance needs updating and revision to be well-adapted to the challenges of evaluating the lawfulness of a given AWS.

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

3.1 **Legal and Procedural Safeguards** We need four kinds of safeguards in this sphere: 1) the legal establishment of clear duties of care and liability mechanisms for each of the primary areas of risk to individual rights, public safety and human security noted in section two; 2) the legal prohibition of (and efforts to gain international cooperation in prohibiting) applications of AWS that pose a clear threat to human rights/IHL and international security (such as LAWS that use learning algorithms to identify, target and kill people without human authorization or veto/recall opportunity); 3) publicly transparent procedures for international and national regulation of AWS and for obtaining legal redress of harm from AWS; 4) clear establishment of independent regimes of testing, auditing, monitoring, evaluation and incident/fault reporting for AWS (even semi-autonomous systems) with clear chains of accountability for compliance with such independent oversight.

3.2 **Technological Safeguards** Interdisciplinary teams of computer scientists, data scientists, ML engineers, ethicists, legal scholars, designers, social scientists, and human-computer interaction experts working together in the fields of Responsible AI, AI Trust and Safety, and AI ethics have developed a large body of existing expertise in the testing, auditing, evaluation, monitoring, assurance, moderation and fine-tuning of machine learning models for safety and trustworthiness. This work is already in use to make AI systems safer and should form the basis for assurance of AWS. This achievement should not be confused with the entirely separate narrative of 'AI Safety' that has emerged over the past year. That narrative is both highly speculative and narrowly technical, in a way that if enacted, would *constrict* our actual safety capabilities. This narrative is being promoted by a convergence of the 'effective altruist' movement with some large tech company leaders and wealthy investors, who benefit from shifting attention away from present risks of autonomous systems already being deployed as part of their core business models, to a longer-term horizon. Those interests have captured political influence for this new narrative over the objections of many computing researchers and professionals around the world, who note that it reduces 'safety' from the robust meaning that traditional safety professionals would understand, to a very narrow and speculative set of long-term concerns about so-

called ‘x-risk’ (existential risk) from AGI. There are indeed reasons to be concerned about the risks of new, more powerful AI models, both in the near-term and the far-term, and safety research ought to be an urgent priority due to the well-known difficulties of predicting and controlling the behaviour of these types of models, as noted elsewhere in this response. Yet the safeguards that we need are not ‘purely technical’ programmes but require leveraging the comprehensive interdisciplinary expertise that has already been built in the fields of trust and safety and responsible AI.

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?

4.1. **Sufficiency of IHL:** The approach taken by defence organizations to the deployment of AI weapons typically emphasizes features such as safety, security, reliability, dependability, and predictability. In contrast, IHL makes broad, principled demands for foresight (predictability), governability and explainability, as prerequisites for being able to review the lawfulness of a capability and as part of the duty to ensure respect for the principles of IHL at all times. Articulating the connection between safety and reliability requirements and the legal context requires additional work. Evaluating the legality of an AI weapon under IHL would require at least:

4.1.2 Foresight sufficient to foresee situations in which the AI weapon could breach prohibitions. For example, that it is impossible to use the weapon in a discriminating way in certain environments, or where certain key parameters change.

4.1.3 The ability to exercise control over the system to assure that rules governing the conduct of hostilities (distinction, proportionality, precaution) are observed.

4.1.4 The ability to provide an adequate explanation of the operation, performance and effects of a system across expected operational contexts, particularly when an AWS is driven by a deep learning model that is intrinsically opaque to human inspection and interpretation.

5. What are your views on the Government's AI Defence Strategy and the policy statement ‘Ambitious, safe, responsible: our approach to the delivery of AI-enabled capability in Defence’? Are these sufficient in

guiding the development and application of AWS? How does UK policy compare to that of other countries?

5.1 **Views on AI Defence Strategy:**

5.1.1 **Risk:** The strategy and policy commit to safety, reliability and responsibility (and commitment to comply with International Humanitarian Law). This will involve the construction of some sort of assurance argument that will be based on some sort of "safety case". The construction of safety cases includes some assessment of the risks associated with the operation of the system. Difficulties around risk analysis propagate and affect most aspects of the wider assurance case that would include data quality, hardware reliability, software quality, integration issues in systems of systems. In the case of AI weapons, two prominent challenges in risk assessment are:

5.1.1.1 The absence of extensive operational experience to provide estimates of the likelihood and severity of an identified adverse event.

5.1.1.2 In the case of AI weapons, there will be concerns to control unlikely, high consequence events. These events pose severe challenges in generating evidence of the accuracy of estimates and this limits confidence in any assurance argument.

5.2. **Sufficiency of guidance in the strategy:** The strategy is ambitious and considers the adoption of current AI technologies and subsequent transformational capabilities predicted for AI weapons. Characterizing the likely future generations of AI weapons is highly uncertain. However, the strategy provides a framework that will require repeated updating as the capabilities of future AI weapons become clearer. The commitment to establishing a clear framework for teams involved in the development, deployment and operation of AI weapons is a sound starting point, but it is unclear how this commitment will be fulfilled and evolved with changes in the capability of AI weapons.

5.3 **International comparison:** NATO's Principles of Responsible Use are aligned broadly speaking with the AI Strategy. It would not be difficult to show high level consensus on values and commitments, but there is variation in formulation and expression. Working internationally towards common standards and governance approaches would be important over time.

6. Are existing legal provisions and regulations which seek to regulate AI and weapons systems sufficient to govern the use of AWS? If not, what reforms are needed nationally and internationally; and what are the barriers to making those reforms?

6.1 **Sufficiency of existing legal/regulatory provisions:** Article 36 of the Geneva convention provides for an overarching approach to review of AWS. However, the dynamic evolution of AWS suggests that a more frequent and iterative approach to review will be essential. The Stockholm International Peace Research Institute's report, Autonomous Weapon Systems and International Humanitarian Law (June 2021) identifies three conditions that are prerequisites for the evaluation of AWSs. Each of the conditions should be secured by the involvement of human agents. These human agents will also need to have specific (professional) capabilities in order to ensure that their involvement is effective. The clarification of the required human roles and their capabilities is one area where further development is certainly necessary. The three conditions are:

6.1.1 The ability reliably to *foresee* whether the effects of an AWS would contravene prohibitions and restrictions on the conduct of hostilities.

6.1.2 The ability to *administer* the operation of an AWS that is consistent with the rules governing the conduct of hostilities.

6.1.3 The ability to *trace* the operation, performance and effects of AWS back to the relevant human agents.

Professor Stuart Anderson
Professor Alex Lascarides,
Professor Subramanian Ramamoorthy
Professor Shannon Vallor
University of Edinburgh
October 2023

References

1. Shannon Vallor and Bhargavi Ganesh, 'Artificial Intelligence and the Imperative of Responsibility: Reconceiving AI Governance as Social Care,' in *The Routledge Handbook of Philosophy of Responsibility*, ed. Max Kiener, in press, 2023.
2. Gary E. Marchant, Braden Allenby, Ronald Arkin, Edward T. Barrett, Jason Borenstein, Lyn M. Gaudet, Orde Kittrie, Patrick Lin, George R. Lucas, Richard O'Meara, Jared Silberman,

'International Governance of Autonomous Military Robots,' *The Columbia Science and Technology Law Review,* Vol, XII 2011, pp. 272-315. https://academiccommons.columbia.edu/doi/10.7916/D8TB1HDW

3. Peter W. Singer, *Wired for War*: *The Robotics Revolution and Conflict in the 21st Century*, London: Penguin, 2009.
4. Michael Skerker, Duncan Purves, Ryan Jenkins. 'Autonomous weapons systems and the moral equality of combatants.' *Ethics and Information Technology* 22, 197–209 (2020). https://doi.org//10.1007/s10676-020-09528-0
5. Shannon Vallor, 'The Future of Military Virtue: Autonomous Systems and Moral Deskilling in the Military Profession.' In *2013 5th International Conference on Cyber Conflict (CyCon 2013): Proceedings*, Karlis Podens, Jan Stinissen and Markus Maybaum, eds. (Tallinn, Estonia: NATO CCDCOE, 2013), pp. 471-486. https://www.ccdcoe.org/uploads/2018/10/9_d2r1s10_vallor.pdf
6. Jennifer Cobbe, Michael Veale, and Jatinder Singh, 'Understanding Accountability in Algorithmic Supply Chains,' *FAccT '23: Proceedings of the 2023 ACM Conference on Fairness, Accountability and Transparency.* June 2023, pp 1186-1197. https://dl.acm.org/doi/10.1145/3593013.3594073
7. ISO/IEC 23894:2023 standard, 'Information technology – Artificial Intelligence - Guidance on Risk Management,' February 2023. https://www.iso.org/standard/77304.html
8. Netta Goussac, Laura Bruun and Vincent Boulanin, 'Autonomous Weapon Systems and International Humanitarian Law.' Stockholm International Peace Research Institute, June 2021. https://www.sipri.org/publications/2021/other-publications/autonomous-weapon-systems-and-international-humanitarian-law-identifying-limits-and-required-type