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

18 November 2020

English only

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System
Geneva, 21–25 September 2020 and 2–6 November 2020
Agenda item 5
Focus of work of the Group of Governmental Experts in 2020, in line with agreed mandate

United Kingdom Expert paper: The human role in autonomous warfare

This paper is intended to generate dialogue and stimulate debate in order to further discussions on the human role in autonomous weapons systems. It does not represent fully formed policy positions and should not be read as a formal representation of UK policy.

I. Overview

- There are clear operational, legal and ethical drivers for focussing on the human role within the use of force.
- It is important to build a shared understanding of the multidimensional nature of human control over weapon systems.
- There is a need to understand the practical activities that occur throughout the lifecycle of a weapon system and how they collectively contribute towards human control over weapon systems and compliance with International Humanitarian Law.
    - What the individual activities look like in practice and how they fit together need to be determined - rather than viewing the activities in isolation, a through-life approach needs to be applied which considers how they interact across the legal, technological and military domains.

II. Introduction

1. The essential role of humans throughout the development, deployment and use of weapon systems with autonomous functions has been an enduring theme throughout the discussions held by the LAWS Group of Governmental Experts (GGE). This led to the adoption of an additional guiding principle within the 2019 report of the GGE focussing specifically on human-machine interaction[1]. This reflects the broad consensus among states

[1] Principle (c): "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-

regarding the centrality of this topic to the debate. The real challenge arises when attempting to reach common understanding on what the human role should look like in practice given the breadth of weapon type and scenarios which could be considered. The concept of human control over the use of force lies at the heart of this debate and is subject to a wide variety of interpretations.

2. Whilst the United Kingdom (UK) does not possess or wish to develop fully autonomous weapon systems operating outside of any form of human control, it recognises that autonomous systems could lead to both military and humanitarian advantages. The UK therefore seeks to embrace the benefits of autonomy, including associated areas of technology like Artificial Intelligence (AI), whilst identifying and mitigating the potential risks, as is the case for any area of emerging technology. Automating some tasks within the targeting process can provide both military and humanitarian benefits, but there is always an essential role for human judgement within this process and throughout the wider lifecycle of a weapon system. Rather than removing the human role entirely, the introduction of autonomous functions within weapon systems changes this role, resulting in potential risks and opportunities.

3. This paper builds on the UK working paper published in 2018[2] by describing the UK's perspective on the human role in weapon systems in more detail - specifically the concept of human control. In doing so it seeks to stimulate discussions and inform the ongoing work of the GGE. First it sets out the grounds for focussing on human control. This is followed by a description of the various dimensions of human control which should be taken into consideration. It addresses two main questions: why is human control over the use of force an important concept within the GGE; and what does this concept mean?

III. The case for Human Control

4. The GGE's enduring focus on the human role in autonomous weapons is well founded. Two key motivators stand out: the pursuit of operational advantage; and compliance with relevant legal and ethical demands.

Operational advantage

5. Appropriate control of military systems provides operational benefits in terms of both optimised performance and risk reduction. Military success is never solely dependent on technological advances. Rather it is the way in which these technologies are used to augment and extend human capabilities that often proves decisive. It is true that both humans and highly automated systems are subject to major limitations and vulnerabilities. For example, people are susceptible to numerous cognitive biases and the effects of fatigue and stress, whilst advanced automation can be brittle in terms of its ability to deal with unusual or ambiguous situations (for example image classification using Deep Neural Networks[3]). It is therefore imperative to seek the right balance of function allocation and interaction between humans and machines in order to avoid undesirable consequences, and in doing so achieve military and humanitarian benefits. The UK describes this perspective as **human-machine teaming**, an approach which recognises that the integration of humans and machines working

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." Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, (2019). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.* Geneva: The United Nations Office at Geneva.

[2] United Kingdom, (2018). *Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles.* Geneva: The United Nations Office at Geneva.

[3] Alcorn, M.A., Li, Q., Gong, Z., Wang, C., Mai, L., Ku, W.-S., and Nguyen, A. (2019). Strike (With) a Pose: Neural Networks Are Easily Fooled by Strange Poses of Familiar Objects. In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach: IEEE, pp. 4840-4849.

towards a common goal, with their relative strengths and weaknesses, is key to military success.[4]

Legal and ethical compliance

6. It is a point of broad agreement within the GGE that the use of lethal force, and the weapon systems used to achieve this, must comply with International Humanitarian Law (IHL). As a signatory to Article 36 of the 1977 Additional Protocol 1 to the Geneva Conventions, the UK conducts legal weapons reviews[5] of all qualifying systems. They are conducted at key milestones in the procurement process to assure the legality of a new weapon, means or method of warfare throughout its study, development, acquisition and adoption.

7. During operations IHL requires that a weapon system must be operated in accordance with four basic principles of distinction, military necessity, humanity, and proportionality[6]. To maintain compliance the same Rules of Engagement (ROE) apply, whether the system is manned or unmanned and regardless of the degree to which it can operate autonomously, meaning that "targets must always be positively identified as legitimate military objectives"[7]. It is clear that humans are obliged to comply with IHL throughout the development and use of weapon systems[8]. However, what this means in practice in terms of human supervision, the predictability and reliability of a weapon system, and other operational constraints[9] is complex and context dependent.

8. Equally the use of force must also be consistent with the ethical standards which underpin this body of law. Two ethical arguments stand out as particularly relevant to LAWS, and both point towards the need for a continued focus on the human role in the use of force; assigning responsibility and preserving dignity.

9. No matter how sophisticated the tools of warfare become, humans as moral agents are ultimately held accountable for the use of force. Legal or moral responsibility cannot be delegated to these aforementioned tools, however intelligent they may be[10]. This is a point which has justifiably been repeated by numerous parties and is stated within the guiding principles of the GGE[11]. A great deal of attention has been focussed on the alleged responsibility gap associated with LAWS. In summary this might entail a weapon system behaving in an unexpected way and, due to its complexity and unpredictability, those involved in its development or deployment neither intended nor foresaw the outcome. Assigning moral responsibility for any resulting harm can therefore be problematic.

[4] Development, Concepts and Doctrine Centre, (2018). *Joint Concept Note 1/18: Human-Machine Teaming.* Swindon: UK Ministry of Defence.

[5] Development, Concepts and Doctrine Centre, (2016) *UK weapon reviews*. Swindon: UK Ministry of Defence.

[6] UK Foreign and Commonwealth Office, (2018). *Guidance: The UK and international humanitarian law 2018.* [online] Available at https://www.gov.uk/government/publications/international-humanitarian-law-and-the-uk-government/uk-and-international-humanitarian-law-2018.

[7] Development, Concepts and Doctrine Centre, (2017). *Joint Doctrine Publication 0-30.2: Unmanned Aircraft Systems.* Swindon: Ministry of Defence.

[8] Davison, N. (2017). "A legal perspective: Autonomous weapon systems under international humanitarian law". In: *UNODA Occasional Papers No. 30, November 2017: Perspectives on Lethal Autonomous Weapon Systems*. New York: United Nations, pp. 5-18.

[9] International Committee of the Red Cross, (2018). *The Element of Human Control*. Geneva: The United Nations Office at Geneva.

[10] The Canberra Working Group, (2019). *Guiding Principles for the Development and Use of LAWS: Version 1.0*. Geneva: The United Nations Office at Geneva.

[11] Principle (b): "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". Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, (2019). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.* Geneva: The United Nations Office at Geneva.

10. In a military context, moral responsibility is vested in the personnel who employ weapon systems and is discharged through the military chain of command as well as command and control (C2) procedures. This does however necessitate that those held responsible have a sufficient understanding of the capabilities and limitations of the weapon system, and of the environment in which it is to be deployed; a point which again highlights the importance of the appropriate form of human-machine interaction along with broader considerations throughout the wider lifecycle such as training of military personnel. These points relating to human-machine interaction are discussed in more detail later in this paper. The avoidance of such a responsibility gap is one of the drivers for the UK position that it will always operate its weapons under human control to ensure accountability for weapon usage. The complexity comes when considering what this control should look like in practice; a topic discussed later in this paper.

11. Dignity is a central notion within many ethical frameworks. The attribution of dignity to all people means they can expect certain treatment at the hands of others; primarily that they should be treated with respect[12]. Even adversaries in armed conflict should be treated with respect in order to maintain the morality of warfare[13]. The relevance of dignity to LAWS centres on the fact that a weapon system is not a moral agent. To treat a human with dignity means, in part, that humans as moral agents must exercise judgement within the military targeting cycle. Importantly, when viewed alongside a commitment to compliance with IHL, preserving dignity in the use of lethal force requires some form of human control during the design and use of a weapon system[14].

12. Finally, the potential legal and ethical advantages of using autonomy to enhance control over a weapon system, for example by augmenting human decision making or extending control within challenging environments or timescales, should not be ignored. These advantages should in fact be pursued if one’s goal is to enhance control over weapon systems, uphold compliance with IHL, and thereby avoid undesirable unintended consequences.

IV. Characterising Human Control

13. Human control is complex, dynamic, multidimensional and situation dependent. Three categories of control measures were proposed in a recent report jointly published by the Stockholm International Peace Research Institute (SIPRI) and the International Committee of the Red Cross (ICRC): the weapon system’s parameters of use, the environment, and human-machine interaction[15]. These clearly highlight the diverse means by which control can be exercised over the use of force and echoes previous work by the ICRC on the various dimensions of human control of weapon systems[16]. Building on this work and others, this section describes the nature of human control over weapon systems in terms of human-machine interaction; the distributed nature of control; the impact of context; and the importance of considering the whole system lifecycle.

[12] Taylor, I. (2020). *Literature Review: Ethical Challenges of AI in Weapons Systems.* London: The Alan Turing Institute.

[13] Nagel, T. (1972). War and Massacre. *Philosophy & Public Affairs,* 1(2), pp. 123-144.

[14] International Panel on the Regulation of Autonomous Weapons, (2018). *Focus on Ethical Implications for a Regulation of LAWS*. Berlin: German Institute for International and Security Affairs.

[15] Boulanin, V., Davison, N., Goussac, N., and Carlsson M.P. (2020). *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*. Solna: Stockholm International Peace Research Institute.

[16] International Committee of the Red Cross, (2018). *The Element of Human Control*. Geneva: The United Nations Office at Geneva.

Control through human-machine interaction

14. When the interaction between humans and autonomous systems is examined more closely, several important dimensions emerge. These are presented in Table 1. Unlike traditional taxonomies which describe autonomy in terms of discrete levels on a scale, each of these dimensions can be viewed as sitting on a continuum. For example, it is unlikely that every aspect of system behaviour would need to be completely predictable to an operator. Instead it is important to consider which system behaviours need to be predictable and to what degree. The appropriate level of control as described by these dimensions is highly situation dependent and is not necessarily bound to a single moment in time.

Table 1. Dimensions of human control: adapted from Boardman and Butcher (2019)[17].

Dimension	*Description*
Freedom of choice	The degree of freedom the human has to choose between possible courses of action. This freedom could be constrained by multiple factors including system design, organisational culture, and workload. It should be noted that in some circumstances it may not be desirable to provide complete freedom of choice due to factors such as the required speed of decision making or workload.
Ability to impact	The human's ability to impact and change the behaviour of the system, either in real time or in advance by setting boundaries or constraints.
Time to decide	Refers to whether a human has sufficient time to process information, make decisions and impact the behaviour of the system if required. There are situations where direct interaction with the system at the time of an effect is neither feasible nor desirable, meaning that constraints on the system's behaviour are set in advance.
Situation understanding	The extent to which a human accurately understands the real world situation. This includes their perception of elements and events within the environment with respect to time and space and the comprehension of their meaning.
System understanding	The extent to which a human accurately understands the system state. This might include the provenance, quality and accuracy of information presented to them and the rationale for decisions or recommendations made by the system.
Predictability	The extent to which the human can accurately project how the system will behave and interact with its environment.

Control is distributed

15. When the way in which humans are currently involved in the military targeting process is examined the highly distributed nature of control, between people and over time, becomes clear. Conventional air operations are a case in point: control is not centralised with

[17] Boardman, M. and Butcher, F. (2019). An Exploration of Maintaining Human Control in AI Enabled Systems and the Challenges of Achieving It. In: *NATO IST-178 Workshop on Big Data Challenge-Situation Awareness and Decision Support*. Brussels: North Atlantic Treaty Organization Science and Technology Organization.

the pilot but instead multiple people exercise different forms of control throughout the targeting process including before the weapon is activated[18]. Within military operations critical decisions about the use of force are taken at various levels of command; tactical, operational and strategic, often well in advance of the deployment of a weapon system[19]. As others have argued, solely relying on an operator making decisions in the heat of the moment as a panacea for human control is never the safest approach[20]. Likewise, anyone familiar with James Reason's Swiss cheese model of accident causation[21] will appreciate that putting all your faith in a single defence against failure, or not considering how multiple layers of defences may interact, is unwise. This last point highlights the importance of clear and comprehensive processes for implementing control measures throughout a weapon lifecycle, like those already presented to the GGE by states including the UK[22] and Australia[23], and discussed at previous meetings of the GGE.

Control is context dependent

16. Context matters when considering appropriate control measures. The nature of the task and the environment should have major implications for how control is implemented. For example, a pre-planned targeting activity against a known objective versus a reactive engagement in self-defence might require different forms of control. Equally the operational environment, including its complexity and time constraints, will also have an impact. For example, as the time available to make decisions decreases the need to rely on control measures enacted earlier in the targeting process is likely to increase.

Control throughout the lifecycle

17. Finally, it is important to consider not just how, but also *when* human control is exercised. The GGE have recognised within their guiding principles that human responsibility for the use of force is not confined to an individual operator but extends across the lifecycle of a weapon system[24]. The GGE have already identified multiple phases throughout a weapon lifecycle where control measures can and must be implemented: political direction; research and development; testing, evaluation and certification; deployment, training, command and control; use and abort; and post-use assessment[25]. Some of the control measures proposed by the SIPRI and the ICRC[26] and touched on in previous

[18] Ekelhof, M. (2019). Moving Beyond Semantics on Autonomous Weapons: Meaningful Human Control in Operation. *Global Policy*, 10(3), pp. 343-348.

[19] United Nations Institute for Disarmament Research, (2020). *The human element in decisions about the use of force*. [online] Available at: https://unidir.org/publication/human-element-decisions-about-use-force.

[20] Lewis, L. (2018). *Redefining Human Control: Lessons from the Battlefield for Autonomous Weapons.* Arlington: CNA Centre for Autonomy and Artificial Intelligence.

[21] Reason, J. (2000). Human error: models and management. *BMJ*, 320, pp. 768-770.

[22] United Kingdom, (2018). *Human Machine Touchpoints: The United Kingdom's perspective on human control over weapon development and targeting cycles.* Geneva: The United Nations Office at Geneva.

[23] Australia, (2019). *Australia's System of Control and applications for Autonomous Weapon Systems*. Geneva: The United Nations Office at Geneva.

[24] Principle (b): "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". Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, (2019). *Report of the 2019 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.* Geneva: The United Nations Office at Geneva.

[25] Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, (2018). *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems.* Geneva: The United Nations Office at Geneva.

[26] Boulanin, V., Davison, N., Goussac, N., and Carlsson M.P. (2020). *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*. Solna: Stockholm International Peace Research Institute.

meetings of the GGE, such as restricting the type of target and task, temporal and spatial constraints, constraining weapon effects, allowing for deactivation and fail-safe mechanisms where appropriate, and controlling the environment to exclude civilians or civilian objects, point towards decisions and activities which must be conducted much earlier in the lifecycle than the point of weapon use. This lifecycle approach demonstrates the various control measures and processes which can exist throughout the design and development, test and evaluation, training and deployment, use, and even after-action evaluation of a weapon system.

V. Conclusion

18. This working paper has addressed two key questions: why is it important for the GGE to continue its focus on the human role within the use of force, and more specifically the concept of human control? And what is meant by the term? In response to the first question; operational, legal and ethical arguments have been described, pointing towards human control as an enabler of military effectiveness and avoidance of undesirable unintended consequences. Second, the concept of human control was described in terms of human-machine interaction, its distributed nature, the impact of context, and the importance of considering the whole system lifecycle. It is anticipated that this dynamic and multidimensional description of human control will contribute towards a common understanding of this central topic within the GGE. An important next step for the GGE will be to build a shared understanding of what this means in terms of legal, technical and military activities throughout a weapon lifecycle.

19. The last point relating to the whole systems lifecycle is particularly important when considering practical approaches for operationalising the GGE guiding principles. Whilst there is no one-size-fits-all solution to the human role within the use of force, the lifecycle approach discussed in previous sessions of the GGE provides a firm foundation for the systematic consideration of control measures. This should serve as a framework for rigorous identification and implementation of good practices in through-life activities such as research and development, design, test and evaluation, legal review, training of personnel, and deployment of weapon systems. In doing so this framework and compendium of good practice would serve as a tool for operationalising the guiding principles at a national level throughout the lifecycle of a weapon system, thereby helping to address the potential opportunities and risks posed by emerging technologies in the area of LAWS. It is only through this kind of multidimensional, process-driven approach that the human role within the use of force can be optimised to enable military effectiveness whilst concurrently pursuing humanitarian goals.