UNIDIR on Lethal Autonomous Weapons

Mapping our Research to the Discussions of the GGE on LAWS

ALICE SPAZIAN
ARTHUR HOLLAND MICHEL
ALISHA ANAND

⊚ **UNIDIR** UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH

UNIDIR **SECURITY** AND **TECHNOLOGY** PROGRAMME

ACKNOWLEDGEMENTS

ABOUT UNIDIR

The United Nations Institute for Disarmament Research (UNIDIR) is a voluntarily funded, autonomous institute within the United Nations. One of the few policy institutes worldwide focusing on disarmament, UNIDIR generates knowledge and promotes dialogue and action on disarmament and security. Based in Geneva, UNIDIR assists the international community to develop the practical, innovative ideas needed to find solutions to critical security problems.

CITATION

A. Spazian, A. Holland Michel, A. Anand, *UNIDIR on Lethal Autonomous Weapons: Mapping our Research to the Discussions of the GGE on LAWS*, Geneva, Switzerland: UNIDIR, 2021.

NOTE

TABLE OF CONTENTS

ABOUT THE AUTHORS

Alice Spazian is a Graduate Professional for the Security & Technology programme. Her work focuses on artificial intelligence, specifically autonomous weapons systems, as well as on the impact of emerging technologies on international security. She holds a Master's degree in International Security from the Sant'Anna School of Advance Studies jointly with the University of Trento. In her previous positions, Alice worked in the human rights office of the municipality of Bologna and as a researcher in the field of interfaith dialogue.

Arthur Holland Michel is an Associate Researcher with the Security and Technology Programme at UNIDIR. He is the founder of the Center for the Study of the Drone at Bard College, where he was Co-Director from 2012 to 2020, and currently serves as a Senior Fellow focusing on autonomy and advanced surveillance technology at the Carnegie Council for Ethics in International Affairs. He has written widely for popular and academic media about unpiloted systems, artificial intelligence and other emerging security and surveillance technologies. His first book, Eyes in the Sky: The Secret Rise of Gorgon Stare and How It Will Watch Us All, was published by Houghton Mifflin Harcourt in 2019. Follow Arthur on Twitter @WriteArthur.

Alisha Anand is the Research Assistant for Security and Technology at UNIDIR. Before joining UNIDIR, she worked on non-proliferation and export controls with the Disarmament and International Security Affairs Division of the Indian Ministry of External Affairs, the Manohar Parrikar Institute for Defence Studies and Analyses and the Federation of Indian Chambers of Commerce & Industry. She holds a Master's degree in Law and Diplomacy from the Fletcher School, Tufts University, where she specialized in international security and international law. Follow Alisha on Twitter @AlishaAnand912.

INTRODUCTION

The United Nations Institute for Disarmament Research has been studying the weaponization of increasingly autonomous technologies since 2013. Over this period, the Institute has produced 15 in-depth studies on this issue.[1] These projects have largely focused on considerations that the Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems (LAWS) has identified as having particular relevance to its development of consensus-based recommendations on a potential normative and operational framework for LAWS. To support the GGE's ongoing discussions in 2021 and beyond, this report summarizes a range of key contributions that these studies have made in these areas. Each section provides a brief summary of the topic as it has been discussed within the context of the Group, followed by a summary of our relevant findings sourced to the reports in which those findings are presented.

1. DEFINING AUTONOMOUS WEAPONS

Setting a definition for autonomous weapons and determining whether and how to apply any such definition is widely viewed within the GGE as an important prerequisite for any firm normative action. The GGE has considered a range of different perspectives concerning both the scope and utility of a possible definition of LAWS.[2] In particular, some parties have sought to establish a definition of 'fully autonomous' weapons that would serve as the scope of a proposed LAWS ban. Such systems would be distinguished from 'partially autonomous' or 'semi-autonomous' systems, which would not be banned (but may be subject to other frameworks).[3] However, many parties have acknowledged the challenges of setting and applying such bright definitional lines. The GGE has also considered the notion of "a technology-neutral approach" to defining autonomy, which would anchor any definitional markers of 'autonomy' to the human element in the use of such systems rather than to the technical characteristics of these weapons.[4]

UNIDIR'S FINDINGS

- Our work has emphasized that 'autonomy' — a system's freedom and latitude in how it achieves its goals—is not a binary characteristic and is not to be confused with 'intelligence', which is a system's ability to determine the best course of action to achieve its goals.[5]

---

1   A full list of these publications is available at the following webpage: https://unidir.org/projects/artificial-intelligence-and-weaponization-increasingly-autonomous-technologies
2   Report of the 2017 GGE on LAWS (Annex II, para. 2).
3   GGE Informal Consultations, 28 June to 2 July 2021; for more on the possibilities and challenges of distinguishing 'fully', 'semi' and 'partially' autonomous systems, see also the 2021 written contributions of the Non-Aligned Movement, Canada, United States, Philippines, France, Kazakhstan, United Kingdom, Poland, and the Russian Federation.
4   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 20).
5   The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence (2018).

- Building on both our exploration of the technical characteristics of LAWS and our close monitoring of the discourse on LAWS over the years, we identified three main approaches to the development of a definition of autonomous weapon systems (AWS): a technology-centric approach focused on systems' technical attributes; a human-centric approach that defines levels of system autonomy according to the human's role in its operations; and a task/function approach based on identifying those actions which the system can execute autonomously.[6] These are not necessarily mutually exclusive perspectives; indeed, it might be possible to develop a comprehensive scope of 'autonomy' through a combination thereof.[7]
- We have also emphasized that it is crucial to consider LAWS holistically, as 'systems of systems' that integrate algorithmic components with sensors, actuators, control interfaces, and many other elements, rather than solely seeking to characterize and define their artificial intelligence (AI) components (i.e., their data processing systems).[8] This approach serves a more grounded and comprehensive consideration of how autonomous systems will behave in the real world.
- This all being said, there may be some limits to the extent to which words can capture certain technical dimensions and gradations of autonomy, especially when it comes to complex algorithmic elements and complex control architectures.[9]

2. WEAPON SYSTEM CHARACTERISTICS

Developing a common understanding of the characteristics of LAWS is also seen as an important precursor for any consensus-based recommendations in the GGE. Such characteristics include the "role and impacts of autonomous functions in the identification, selection or engagement of a target";[10] LAWS's capacity for "self-learning and evolution" or the ability to "self-initiate";[11] the "nature of communications links with human operator or chain of command; whether intervention in the system's operation would be possible once activated; reliability and predictability"[12] and system vulnerabilities;[13] "biases in the data sets used to train the machine";[14] the "ability [of militaries] to

---

6   The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics, and Definitional Approaches (2017).

7   Parties might, for example, first identify the role of the human operator, then "identify the tasks where autonomy might call these roles into questions", and finally "describe what technological features/characteristics they would need to see/avoid in future weapon systems". The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics, and Definitional Approaches (2017).

8   Known Unknowns: Data Issues and Military Autonomous Systems (2021).

9   Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems (2020).

10   Report of the 2019 GGE on LAWS (para. 19(a)).

11   Report of the 2019 GGE on LAWS (para. 20(b)); 2020 Session of the GGE on LAWS: Chairperson's Summary (para. 23).

12   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 23).

13   Report of the 2017 GGE on LAWS.

14   Report of the 2019 GGE on LAWS (para. 20); 2020 Session of the GGE on LAWS: Chairperson's Summary (para. 22).

impose operational constraints on a weapon-system";[15] and system features that benefit military commanders' decision making.[16] High Contracting Parties have also deemed human operator understanding of LAWS system characteristics to be a key "to exercise appropriate judgement and ensure that the use of weapon systems is consistently within applicable international law".[17]

UNIDIR'S FINDINGS

- There are many technical factors that limit or augment system intelligence and autonomy, including the system's capacity for goal-satisfying actions, its level of predictability, communication modalities, depth of 'reasoning' or 'problem-solving', the power and precision of system sensors and capacity for information synthesis, as well as the constraints or freedoms that operators may impose on the system's operating location, environment, payloads and functions.[18]

- Intelligent systems do not spontaneously become more autonomous—humans choose to give machines more autonomy as they become more intelligent.[19]

- Despite recent advances in the field, state-of-the-art AI systems today still have significant limitations and vulnerabilities and can only perform narrow tasks under specific conditions. A wide range of errors or unexpected autonomous system actions or effects can arise in *all* kinds of autonomous or semi-autonomous systems.

- Crucially, all autonomous systems are vulnerable to data issues.[20] System testing and risk assessment cannot necessarily validate systems against all the potential issues that a complex autonomous system might encounter in a complex environment.[21]

- Hidden interactions both within the system-of-systems itself, as well as between the system and other systems or agents (especially other autonomous systems, as within a swarm[22]), can create feedback loops or other emergent effects that create unintentional risks. Such phenomena may be more likely to arise in systems reliant on machine learning.[23]

- Machine learning-based systems can exhibit – and fail as a result of – a range of different types of system biases that result from inappropriate system training, coding, design or deployment.[24]

15   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 23).
16   Report of the 2018 GGE on LAWS (para. 24).
17   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 32).
18   Framing Discussions on the Weaponization of Increasingly Autonomous Technologies (2014).
19   The Weaponization of Increasingly Autonomous Technologies: Artificial Intelligence (2018).
20   The most common of which include incomplete data, low-quality data, incorrect or false data, and discrepancies between the data of a system's development and that of its deployment.
21   Known Unknowns: Data Issues and Military Autonomous Systems (2021).
22   Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems (2020).
23   Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016); The Black Box, Unlocked (2020).
24   Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies (2018).

- Technical measures to mitigate these issues may not be fully effective[25] or may introduce new risks.[26]
- System 'predictability' and 'understandability', which are widely seen as vital for autonomous weapons,[27] are not static or universally measurable qualities. Furthermore, they are challenging characteristics to achieve, given the inherently unpredictable and uninterpretable character of autonomous weapons and the complex, dynamic nature of conflict.[28]
- As a result, systems that perform all steps of mission execution autonomously are still, in all likelihood, neither technically feasible, militarily desirable nor legally permissible. On the other hand, we have found that systems that introduce some *constrained* degrees of autonomy in *certain specific tasks under certain specific conditions* could be feasible and legally compliant, provided that legal frameworks are strictly applied to the entire life cycle of the weapon.[29]

3. THE HUMAN ELEMENT

There is a growing consensus within the GGE that "the use of force must reflect human agency and human intention".[30] High Contracting Parties have also noted that human judgement is an essential element to ensure compliance with international humanitarian law (IHL), as it would represent a safeguard for responsibility and accountability.[31] This being said, the types and degrees of human involvement necessary for compliance are yet to be defined. Parties have recognized that human–machine interaction may take many forms depending on the system's life cycle, as well as its characteristics and the operational context. Some parties have also noted that "effective human control, involvement or judgment may not necessarily equate to direct, manual control"[32] and that appropriate types and levels of control vary in different contexts. Consensus is also emerging around the notion that control should be exercised, at a minimum, on any system's lethal functions.[33]

---

25   For example, while it is thought that "algorithms expressed in software code that can be inspected and altered by the human designers of the system in question" constitute a robust safeguard, it is not clear yet to what extent this safeguard can effectively prevent hidden interactions in highly autonomous systems, particularly those employing self-optimization routines. Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).

26   For instance, the redundancy measures that might be introduced to mitigate such risks could, in some cases, increase system complexity or undermine effective human operator oversight (a phenomenon known as automation bias). Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).

27   The Black Box, Unlocked (2020).

28   The degree to which operators can understand and anticipate systems depends on a wide range of external factors, including the actors' capacity for understanding the system and the characteristics of the system environment. The Black Box, Unlocked (2020).

29   Table-Top Exercises on the Human Element and Autonomous Weapons Systems: Summary Report (2021).

30   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 28).

31   Report of the 2019 GGE on LAWS.

32   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 30).

33   Report of the 2017 GGE on LAWS (para. 4).

UNIDIR'S FINDINGS

- The human element in the use of force is operationalized by a chain of actors[34] over the course of a process that often begins long before the attack itself begins.[35] Similarly, human–machine interaction might be enacted through a range of different temporal, technical, and organizational modalities. The human element is thus "distributed and context dependent".[36]
- Decisions on the appropriate type and level of human control in any given operation will be informed by:
  » the technical characteristics of the system;
  » the doctrine, concept of operations, specific rules of engagement and restrictions/guidelines that were developed for the system during testing and review;
  » the characteristics of the operation, the environment, the target and the adversary;
  » the military advantage/disadvantage of employing autonomous functions in that context; and crucially
  » any relevant legal considerations.[37]
- In any context and arrangement, successfully implementing the human element depends on achieving appropriate system understandability and predictability,[38] as well as measures to properly mitigate unintentional risks.[39]

4. RESPONSIBILITY AND ACCOUNTABILITY

The GGE on LAWS has affirmed in its Guiding Principles that "human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines." These Principles also affirm that "accountability for developing, deploying and using any emerging weapon system in the framework of the CCW 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."[40] Building on this shared view, parties have emphasized that effective responses to risks posed by autonomous weapons systems require careful consideration of *how* human

---

34   Namely political leadership, technology developers, system operators, military commanders and advisers, among others.
35   The Human Element in Decisions About the Use of Force (2020).
36   For example, in situations of deliberate targeting, where more time is available to plan an attack, control architectures would be different from situations of dynamic targeting. The Human Element in Decisions About the Use of Force (2020); Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems (2020).
37   Table-Top Exercises on the Human Element and Autonomous Weapons Systems: Summary Report (2021).
38   The Black Box, Unlocked (2020).
39   Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).
40   Report of the 2019 GGE on LAWS (Annex IV, Guiding principle (b) and (d)).

responsibility and accountability can be ensured,[41] and that the application of responsibility and accountability must be a consideration across the entire life cycle of the weapon system.[42]

UNIDIR'S FINDINGS

- It is broadly agreed that delegating an action to an autonomous machine does not absolve the human operator of that system from responsibility. However, if an AWS must make many complex 'decisions' in order to execute a human's commands there could be less direct causality between the time, type, context and intent of the human's order and the effects of the weapon's actions.[43]
- Not only do forms of responsibility apply at all stages of the development and acquisition of systems, but many critical decisions related to the use of force take place well in advance of the moment a weapon system, autonomous or otherwise, is activated in an attack.[44] Others take place long after the deployment of the system.[45]
- Autonomous system characteristics may undermine the capacity for commanders to achieve appropriate trust in the relevant systems and make appropriate decisions on the basis of that trust.[46] Furthermore, responsibility for mitigating or preventing the unintended consequences of an autonomous system's employment does not necessarily rest with a single actor or entity.[47]
- As the complexity of the system and the complexity of the environment in which it is used rises, it becomes inherently more challenging for users at all of these stages to understand the system and anticipate its actions. Requiring actors to understand and anticipate such systems may therefore actually give rise to a higher burden of responsibility for weapons effects compared to the use of conventional weapons.[48]
- It may not be possible to hold actors responsible for harms arising from technical issues in AWS that could not specifically be anticipated. And yet the fact that all autonomous systems fail[49] may itself constitute a basis for anticipating harm, even if one does not know when or how that harm will happen. This 'known unknowns' paradox may pose a significant challenge to the application of responsibility and accountability.[50]

---

41   2020 Session of the GGE on LAWS: Chairperson's Summary (para. 9).
42   Report of the 2017 GGE on LAWS (para. 47); Report of the 2019 GGE on LAWS (Annex IV, Guiding principle (b)); Report of the 2019 GGE on LAWS (para. 12).
43   The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches (2017).
44   The Human Element in Decisions About the Use of Force (2020).
45   For example, in post-use assessment, it is necessary to understand why the autonomous system behaved as it did, understand why certain effects resulted from its employment, and to consider whether those responsible for an autonomous system's use reasonably anticipated the effects of their actions, took reasonable steps to avoid undue harms and followed IHL and rules of engagement in other regards. The Black Box, Unlocked (2020).
46   The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches (2017); Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies (2018).
47   Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies (2018).
48   Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).
49   Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).
50   Known Unknowns: Data Issues and Military Autonomous Systems (2021).

5. MORAL AND ETHICAL CONSIDERATIONS

High Contracting Parties have affirmed that "relevant ethical perspectives, should guide the continued work of the Group".[51] These ethical considerations are derived from the principles of humanity and the dictates of public conscience, as enshrined in the Martens clause and human rights law.[52] Delegations have, for example, stressed moral and ethical concerns related to delegating the decision to use force and make life and death decisions to a machine, given that "robots cannot be moral agents".[53] As a result, proposed recommendations and guidelines relating to human control and involvement, human judgement, human intention and human responsibility in the use of LAWS are not just based on legal requirements, but also on fundamental ethical principles.[54]

UNIDIR'S FINDINGS

- Since the beginning of our work on this issue, we have asserted that ethics and social values stand as a fundamental consideration in the policy debate on LAWS. As we have noted, "consideration of ethical issues takes us beyond whether a given weapon is legally acceptable (can we use it?) to whether it is morally acceptable (should we use it?)". All such matters are key to building a shared understanding of "what's at stake".[55]
- Beyond the aforementioned questions related to human agency, we have emphasized that LAWS have significant implications for "security, the responsible use of force, protection of civilians, and human dignity itself". Conversely, we have noted that there could be, in other contexts, a "moral obligation" to use a weapon if it can reduce overall harm. Questions persist as to how to weigh such potential benefits against these implications.[56]
- We have also considered how the acquisition of advanced weaponized autonomous systems and other forms of military AI by some States may encourage others to follow suit in the pursuit of a 'first mover advantage'. The resulting unchecked proliferation could stand at odds with the rigorous application of ethical principles.[57]

---

51   Report of the 2019 GGE on LAWS (Annex IV, Guiding principles preamble).
52   Report of the 2018 GGE on LAWS.
53   Report of the 2017 GGE on LAWS (para. 8, para. 33).
54   2020 Session of the GGE on LAWS: Chairperson's Summary
55   The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values (2015).
56   The Weaponization of Increasingly Autonomous Technologies: Considering Ethics and Social Values (2015).
57   Modernizing Arms Control: Exploring Responses to the Use of AI in Military Decision-Making (2020).

6. LEGAL REVIEWS

High Contracting Parties have affirmed that "legal reviews of new weapons are essential to ensuring that emerging technologies in the area of LAWS can be used in compliance with applicable rules of international law"[58] but have noted significant challenges that must be addressed in order to enable effective reviews of LAWS.[59] Delegations have considered a range of options for bolstering the legal review process for LAWS, including by way of formal information-sharing, the exchange of best practices,[60] and the establishment of a "regular evaluation process based on a set of criteria relevant for LAWS and applied across the life-cycle of a weapon system".[61]

UNIDIR'S FINDINGS

- To achieve its purpose of ensuring that LAWS can be used in compliance with IHL, a legal review of such a system would have to:
  » anticipate the system's performance in the functions and environments for which it is proposed— particularly by measuring and validating the reliability of the system—and identify any factors that can cause failures (including adversarial data threats or the risk of 'emergent effects');
  » model the ways in which systems could fail and the effects of those failures;[62]
  » determine whether the system's training and testing environments closely match the proposed operational environments and ensure that they do not exhibit harmful forms of bias; and
  » evaluate the degree to which the human element in the planned uses of the system could reliably anticipate or respond to issues that arise.[63]
- Considerable further work is needed to make such reviews possible. Above all, existing test and evaluation techniques for complex autonomous systems will need to be significantly overhauled.[64]
- A precise assessment of whether a system will exhibit the same behaviours in the field that it exhibited in testing will depend on high degrees of predictability and understandability. Achieving such predictability and understandability, and developing metrics and standards to validate them, remains an open technical challenge.[65]

58　2020 Session of the GGE on LAWS: Chairperson's Summary (para. 11).
59　Some of these derive from the secrecy and state-specific practices of legal reviews, leading to a lack of uniformity in the conduct of this process. Other limitations directly originate from "the unique difficulties posed by LAWS for legal weapons reviews, including testing and whether a self-learning system would need to be reviewed every time it changed itself". Report of the 2018 GGE on LAWS (para. 32).
60　Report of the 2017 GGE on LAWS (para. 23); Report of the 2019 GGE on LAWS (Annex III, para. 5).
61　2020 Session of the GGE on LAWS: Chairperson's Summary (para. 11).
62　The Black Box, Unlocked (2020).
63　Known Unknowns: Data Issues and Military Autonomous Systems (2021).
64　Existing testing techniques are ill-suited to identify autonomous system vulnerabilities, rate their likelihood, quantify autonomous system reliability and ensure that the mode of human–machine interaction would enable operators to respond to issues appropriately. Known Unknowns: Data Issues and Military Autonomous Systems (2021); Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies (2016).
65　The Black Box, Unlocked (2020).

- If an autonomous system relies on machine-learning elements that change over time, the initial findings of a review may also require regular revision.[66]
- Because such challenges may make it hard for States to anticipate specific risks in LAWS prior to employment, they could create ambiguity as to human responsibility for the unintended harm caused by autonomous systems once those systems are deployed.[67]

66 Table-Top Exercises on the Human Element and Autonomous Weapons Systems: Summary Report (2021).

67 If, as a result of the limitations of testing and system predictability and understandability, the review and risk assessment process for an autonomous weapon has given a military organization no specific reason to believe that relevant data issues could emerge or result in unintended harm, they would not be required to take those specific data issues into account, or anticipate, preemptively mitigate or respond to those issues during that mission. In the event of harm arising from such 'unknown' failures, these actors could not reasonably be accused of being responsible for not taking the risk of such failures into account. Known Unknowns: Data Issues and Military Autonomous Systems (2021).

The United Nations Institute for Disarmament Research has been studying the weaponization of increasingly autonomous technologies since 2013. Over this period, the Institute has produced 15 in-depth studies on this issue. These projects have largely focused on considerations that the Group of Governmental Experts (GGE) on emerging technologies in the area of lethal autonomous weapons systems (LAWS) has identified as having particular relevance to its development of consensus-based recommendations on a potential normative and operational framework for LAWS. To support the GGE's ongoing discussions in 2021 and beyond, this report summarizes a range of key contributions that these studies have made in these areas. Each section provides a brief summary of the topic as it has been discussed within the context of the Group, followed by a summary of our relevant findings sourced to the reports in which those findings are presented.

UNIDIR UNITED NATIONS INSTITUTE FOR DISARMAMENT RESEARCH