INTERNATIONAL HUMANITARIAN LAW

DOI: https://doi.org/10.24833/0869-0049-2020-2-53-64

Research article
Received May 2, 2020
Approved June 14, 2020

Dustin LEWIS
Harvard Law School Program on International Law and Armed Conflict
175J, Massachusetts Avenue Langdell, Cambridge, MA, United States of America, 02138
dlewis@law.harvard.edu
ORCID: 0000-0003-3843-7665

INTERNATIONAL LEGAL REGULATION OF THE EMPLOYMENT OF ARTIFICIAL-INTELLIGENCE-RELATED TECHNOLOGIES IN ARMED CONFLICT

INTRODUCTION. *In recent years, increased attention has been dedicated at the international level to legal issues concerning the possible employment of artificial-intelligence-related technologies in hostilities in armed conflict. Most prominently, discussions in the framework of the Convention on Certain Conventional Weapons (CCW) have addressed juridical aspects relative to emerging technologies in the area of lethal autonomous weapons systems.*

MATERIALS AND METHODS. *I analyze contemporary intergovernmental debates in the context of the CCW, international legal frameworks pertaining to armed conflict, and developments in relevant technologies. I do so to trace current trajectories and generate an analytical framework to help apply legal responsibility.*

RESEARCH RESULTS. *A disagreement has arisen among certain States in the context of the CCW as to whether to develop a new primary legal norm or whether existing international humanitarian law is sufficient. Taking account of that current normative impasse, I propose an analytical framework aimed at ensuring the applicability of international legal responsibility in respect of the employment of AI-related technologies in armed conflict.*

DISCUSSION AND CONCLUSIONS. *Given the range of relevant technologies, the employment of AI in armed conflict may occur across diverse thematic and functional areas: not only in the conduct of hostilities, including weapons, but also detention, humanitarian services, maritime systems, and many other areas. Identification of the general concepts and specific attributes necessary to apply international legal responsibility across the array of implicated areas may help provide a framework through which to respect the law, guide behavior, pursue accountability, and generate areas of greater normative consensus.*

KEYWORDS: *artificial intelligence, armed conflict, international humanitarian law, international criminal law, state responsibility, war crimes, weapons, means and methods of warfare, detention, warships*

FOR CITATION: Lewis D. International legal regulation of the employment of artificial-intelligence-related technologies in armed conflict. –*Moscow Journal of International Law.* 2020. No. 2. P. 53–64. DOI: https://doi.org/10.24833/0869-0049-2020-2-53-64

The author declares the absence of conflict of interest.

МЕЖДУНАРОДНОЕ ГУМАНИТАРНОЕ ПРАВО

DOI: https://doi.org/10.24833/0869-0049-2020-2-53-64

Исследовательская статья
Поступила в редакцию: 02.05.2020
Принята к публикации: 14.06.2020

Дастин ЛЬЮИС
Harvard Law School Program on International Law and Armed Conflict
Massachusetts Avenue Langdell, 175J, Cambridge, MA, 02138, United States of America,
dlewis@law.harvard.edu
ORCID: 0000-0003-3843-7665

МЕЖДУНАРОДНО-ПРАВОВОЕ РЕГУЛИРОВАНИЕ ИСПОЛЬЗОВАНИЯ ТЕХНОЛОГИЙ, СВЯЗАННЫХ С ИСКУССТВЕННЫМ ИНТЕЛЛЕКТОМ, В ВООРУЖЕННОМ КОНФЛИКТЕ

ВВЕДЕНИЕ. *В последние годы юридические вопросы, касающиеся возможного использования технологий, связанных с искусственным интеллектом, в военных действиях в ходе вооруженного конфликта, стали предметом повышенного внимания на международном уровне. Наиболее масштабные дискуссии по правовым аспектам, связанным с новыми технологиями в области смертоносных автономных систем вооружений, ведутся в рамках Конвенции о запрещении или ограничении применения конкретных видов обычного оружия, которые могут считаться наносящими чрезмерные повреждения или имеющими неизбирательное действие (КНО).*

МАТЕРИАЛЫ И МЕТОДЫ. *Статья посвящена актуальным межправительственным дебатам в рамках КНО, международно-правовым рамкам, относящимся к вооруженному конфликту, развитию соответствующих технологий. Автор прослеживает текущие направления дискуссий и формирует аналитическую основу для содействия применению юридической ответственности.*

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ. *Государства-участники КНО придерживаются различных позиций по вопросу о том, имеется ли потребность в принципиально новом правовом регулировании или же существующего международного гуманитарного права достаточно. Принимая во внимание эти непримиримые разногласия, автор предлагает аналитические рамки, направленные на обеспечение применимости международно-правовой ответственности в отношении использования в вооруженном конфликте технологий, связанных с искусственным интеллектом.*

ОБСУЖДЕНИЕ И ВЫВОДЫ. *Учитывая спектр упомянутых технологий, использование искусственного интеллекта в вооруженном конфликте может осуществляться в различных тематических и функциональных областях: не только в ходе ведения военных действий, включая применение оружия, но и при задержании, оказании гуманитарного содействия, в военно-морских системах, а также многих других сферах. Идентификация общих концепций и специфических характеристик, необходимых для реализации международно-правовой ответственности применительно к спектру таких областей может помочь выработке схем, обеспечивающих уважение права, регулирование поведения, реализацию ответственности, а также расширению потенциально консенсусных сфер правового регулирования.*

КЛЮЧЕВЫЕ СЛОВА: *искусственный интеллект, вооруженный конфликт, международное гуманитарное право, международное уголовное*

право, ответственность государств, военные преступления, вооружения, средства и методы ведения войны, задержание, военные корабли

ДЛЯ ЦИТИРОВАНИЯ: Льиюс Д. 2020.Международно-правовое регулирование использования технологий, связанных с искусственным интеллектом, в вооруженном конфликте. – *Московский журнал международного права.* № 2. С. 53–64. DOI: https://doi.org/10.24833/0869-0049-2020-2-53-64

Автор заявляет об отсутствии конфликта интересов

1. Introduction

Especially since 2014, dozens of States have dedicated significant attention to certain legal aspects concerning the possible employment of artificial intelligence (AI) and related technologies in the conduct of hostilities in armed conflict. In particular, in the context of the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or Have Indiscriminate Effects (CCW), numerous States — as well as members of civil society, academia, and other sectors — have exchanged views on diverse aspects of the possible design, development, testing, ex-ante review, fielding, employment, post-hoc assessment, transfer, and imposition of responsibility relative to "emerging technologies in the area of lethal autonomous weapons systems"[1]. Views on these matters have been expressed in other intergovernmental fora as well, including in the United Nations General Assembly First Committee. Hundreds of commentaries have been published in international scholarly discourse [Lewis, Modirzadeh, Blum 2016:106–128]. Moreover, many legal and policy proposals and analyses have been put forward, not only by States but also by other actors, including non-governmental organizations.

In this article, I seek to propose an analytical framework aimed at the identification of the general concepts and specific attributes necessary to apply international legal responsibility in this area. First, to set the stage, I trace the current status of the CCW's Group of Governmental Experts on Lethal Autonomous Weapons Systems. I also put forward contextual factors that may help explain the increased attention to AI in armed conflict as well as a central disagreement that has arisen around it (section 2). Second, I submit a framework aimed at ensuring international legal responsibility in respect of employments of AI-related technologies in armed conflict (section 3). Finally, I conclude by inviting international actors to elaborate their understandings of the general concepts and specific attributes necessary to apply responsibility in this area.

For this article, I assume a wide definition of artificial intelligence-related technologies drawn from AI science broadly construed [Sayler 2019:1–4][2]. The research underlying this article was undertaken in the context of a research and policy project of the Harvard Law School Program on International Law and Armed Conflict entitled "International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns"[3].

2. Current trajectories concerning the CCW's Group of Governmental Experts

In 2016, following informal meetings of experts over a few years, the Fifth Review Conference of the High Contracting Parties to the CCW established a Group of Governmental Experts (GGE) on Emerg-

[1] Meeting of the High Contracting Parties to the 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. Final report. December 13, 2019 (hereafter CCW 2019). URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/4F3F92951E0022D9C12584F50034C2F4/$file/CCW+MSP+2019+9.pdf (accessed 12.01.2020).

[2] See also: Lewis D.A. *Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider. – ICRC Humanitarian Law and Policy Blog.* March 21, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/ (accessed 29.04.2020).

[3] See: Harvard Law School Program on International Law and Armed Conflict: *"International Legal and Policy Dimensions of War Algorithms: Enduring and Emerging Concerns".* URL: https://pilac.law.harvard.edu/international-legal-and-policy-dimensions-of-war-algorithms (accessed 29.04.2020).

ing Technologies in the Area of Lethal Autonomous Weapons Systems. The GGE held meetings in 2017, 2018, and 2019. The 2019 Meeting of High Contracting Parties to the CCW agreed, by consensus, on a multipart mandate for the GGE for the 2020–2021 period. The GGE is scheduled to meet for a total of ten days in 2020 and between ten to twenty days in 2021, to be decided by the Meeting of the High Contracting Parties in 2020. (At the time of writing, it is uncertain whether the coronavirus pandemic may impact these scheduled meetings.) Under the agreement of the High Contracting Parties to the CCW, the GGE's recommendations will be reported, as appropriate, for consideration at the 2020 Meeting of High Contracting Parties and the 2021 Sixth Review Conference[4].

a. Context

The GGE operates in a context marked by the following five sets of factors.

First, developments have arisen in certain technical fields resulting in new or amplified technological capabilities that might be employed in armed conflict. Those developments include increases in the efficiency of algorithms, computing power, sensor capacity, and the volume and range of available data. Another set of developments relates to an increase in the potential physical distance or amount of time (or both) that constructed systems with partial or full automatic or autonomous navigation may be able to travel at sea, on land, in air, or in outer space. Furthermore, developments in the fields of miniaturization of constructed systems and of interactive capacities between humans and machines (sometimes called "human-machine teaming") and among machines (sometimes referred to as "swarming") may be relevant as well [Ekelhof, Persi Paoli 2020:9–56].

Second, regarding military operations, some armed forces are increasingly relying on combinations of algorithmic, computational, and other data-driven tools and techniques [Kania 2020:2–7; Sayler 2019:9–36; Boulanin, Verbruggen 2017:19–112]. That increased reliance turns in part on perceptions that those technological developments could facilitate military advantages, including increases in speed, accuracy, and economy of resources and decreases in the number of personnel placed at risk of physical harm. Relevant technologies have already been employed in respect of the following sets of tasks:

- The formulation, collection, and evaluation of intelligence involved in military operations;
- The formulation, identification, nomination, and prioritization of targets involved in attacks;
- The choice and selection of weapons, means, and methods of warfare; and
- The making of certain evaluative decisions and normative judgments involved in compliance with legal rules pertaining to the conduct of hostilities, including concerning binding IHL provisions on distinction, proportionality, and precautions in attack [Ekelhof 2018: 14–28][5].

Third, the GGE mandate concerns weapons specifically and, in certain respects, the conduct of hostilities more broadly. However, the potential employment of AI techniques and methods in situations of armed conflict may impact several other areas as well. These areas include detention[6], humanitarian services[7], uninhabited military maritime systems [Nasu, Letts 2020:83–97], and legal advice[8].

Fourth, the discussions in the GGE have arisen against several background elements, including the following:

- Different conceptions among international actors regarding what constitutes the exercise of sufficient human agency, control, judgment, or some combination thereof with respect to particular forms of military operations involving relevant technologies;
- The sheer complexity of existing and emerging technologies in this area, as well as the rapid pace and scale of developments; and

[4] CCW 2019. P. 5.

[5] See also: Ekelhof M.A.C. The Distributed Conduct of War: Reframing Debates on Autonomous Weapons, Human Control and Legal Compliance in Targeting. Diss. Ph.D. Candidate. Amsterdam. 2019. P. 127-225.

[6] See: Bridgeman T. The viability of data-reliant predictive systems in armed conflict detention. – *ICRC Humanitarian Law and Policy Blog*. April 8, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention/ (accessed 29.04.2020); Deeks A. *Detaining by algorithm*. – *ICRC Humanitarian Law and Policy Blog*. March 25, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/03/25/detaining-by-algorithm/ (accessed 29.04.2020).

[7] See: Lewis D.A. AI and Machine Learning Symposium: Why Detention, Humanitarian Services, Maritime Systems, and Legal Advice Merit Greater Attention. – *Opinio Juris Blog*. April 28, 2020. URL: http://opiniojuris.org/2020/04/28/ai-and-machine-learning-symposium-ai-in-armed-conflict-why-detention-humanitarian-services-maritime-systems-and-legal-advice-merit-greater-attention/ (accessed 29.04.2020).

[8] Ibid.

- Secrecy and a lack of trust among certain (sets of) States concerning their respective military-technological capabilities in this area.

Fifth, there is a lack of an agreed definition or set of characteristics relative to "emerging technologies in the area of lethal autonomous weapons systems." States and other participants in the GGE have put forward definitions that vary widely[9].

b. Mandate

One part of the GGE's mandate during the 2020–2021 period concerns the exploration and agreement on possible recommendations for options related to emerging technologies in the area of lethal autonomous weapons systems, such as for potential challenges to IHL[10]. Included among the categories of possible options raised for addressing the humanitarian and international security challenges posed in this area are a legally binding instrument, a political declaration, and clarity on the implementation of existing obligations under international law, in particular IHL[11]. Another part of the GGE's current mandate pertains to the formulation of consensus recommendations concerning the clarification, consideration, and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems[12].

c. Areas of normative consensus and disagreement

To date, the GGE has obtained at least a measure of normative consensus, in particular as reflected in the elaboration of certain guiding principles. Included among them are the following principles: IHL continues to apply fully to all weapons systems, including the potential development and use of lethal

[9] Statement by India to the GGE "Characterisation of the Systems under consideration in order to promote a common understanding on Concepts and Characteristics relevant to the objectives and purposes of the Convention". March 25, 2019. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/F8C1F0AEE961CA93C12583CC00353A09/$file/25+March+2019+-+5(d).pdf (accessed 29.04.2020); Statement by the International Committee of the Red Cross to the GGE. March 25, 2019. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/7EA110E50853887DC12583CC003032E7/$file/ICRC+GGE+LAWS+ICRC+statement+agenda+item+5c+25+03+2019.pdf (accessed 29.04.2020); Proposal by Ireland "Working Definition". August 29, 2018. URL: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/29August_Proposal_Definition_Ireland.pdf (accessed 29.04.2020); Submission by Estonia and Finland to the GGE "Categorizing lethal autonomous weapons systems - A technical and legal perspective to understanding LAWS". August 24, 2018. URL: https://undocs.org/CCW/GGE.2/2018/WP.2 (accessed 29.04.2020); Submission by China to the GGE "Position Paper". April 11, 2018. URL: https://undocs.org/en/CCW/GGE.1/2018/WP.7 (accessed 29.04.2020); Statement Submitted by Germany to the GGE "On a Working Definition of LAWS /"Definition of Systems under Consideration". April 9–13, 2018, URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/2440CD1922B86091C12582720057898F/$file/2018_LAWS6a_Germany.pdf (accessed 29.04.2020); Submission by the Russian Federation to the GGE "Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapons Systems in the Context of the Purposes and Objectives of the Convention". April 4, 2018. URL: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/GGE.1-WP6-English.pdf (accessed 29.04.2020); Submission by the United States to the GGE "Characteristics of Lethal Autonomous Weapons Systems". November 10, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.7 (accessed 29.04.2020); Submission by Belgium to the GGE "Towards a Definition of Lethal Autonomous Weapons Systems". November 7, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.3 (accessed 29.04.2020); Submission of the Netherlands to the GGE "Examination of various dimensions of emerging technologies in the area of lethal autonomous weapons systems, in the context of the objectives and purposes of the Convention". October 9, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.2 (accessed 29.04.2020); Statement by Switzerland to the Meeting of Experts of the Convention on Certain Conventional Weapons "A purpose-oriented working definition for autonomous weapons systems". April 12, 2016. URL:https://www.unog.ch/80256EDD006B8954/(httpAssets)/A204A142AD3E3E29C1257F9B004FB74B/$file/2016.04.12+LAWS+Definitions_as+read.pdf (accessed 29.04.2020); Statement submitted by France to the CCW Meeting of Experts on Lethal Autonomous Weapons Systems "Non Paper: Characterization of a LAWS". April 11–16, 2016. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/5FD844883B46FEACC1257F8F00401FF6/$file/2016_LAWSMX_CountryPaper_France+CharacterizationofaLAWS.pdf (accessed 29.04.2020); Statement of Italy to the CCW Meeting of Experts on Lethal Autonomous Weapons Systems "Towards a working definition of LAWS". April 11–15, 2016. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/06A06080E6633257C1257F9B002BA3B9/$file/2016_LAWS_MX_towardsaworkingdefinition_statements_Italy.pdf (accessed 29.04.2020); Statement of the United Kingdom to the Informal Meeting of Experts on Lethal Autonomous Weapons Systems. April 11–15, 2016. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/44E4700A0A8CED0EC1257F940053FE3B/$file/2016_LAWS+MX_Towardaworkingdefinition_Statements_United+Kindgom.pdf (accessed 29.04.2020).

[10] CCW 2019. P. 5.

[11] Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System: of the 2019 session "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". September 25, 2019. P. 7. (hereafter, GGE 2019). URL: https://undocs.org/en/CCW/GGE.1/2019/3 (accessed 29.04.2020).

[12] CCW 2019. P. 5.

autonomous weapons systems; human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines; and accountability for developing, deploying, and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law[13]. The IHL provisions that most commonly arise in these debates concern distinction, proportionality, and precautions in attack [Singer 2018:213–499; Lewis, Modirzadeh, Blum 2016:64–76]. Focus has also been placed on reviews of weapons, means and methods of warfare[14].

However, disagreements have also arisen. The most prominent divergence concerns whether existing IHL is sufficient to address the range of issues that may arise in this area or whether a new norm needs to be elaborated (and, if so, what the content of that norm should be and what form(s) it should take). For example, calls have been made for the elaboration of a new legal norm aimed at the regulation, prohibition, or some combination thereof of lethal autonomous weapons systems or at least certain weapons involving an autonomous attribute in the "critical functions" of selection of targets and engagement in attacks[15]. In 2018, for instance, Austria, Brazil, and Chile proposed that the High Contracting Parties to the CCW should decide "to establish an open-ended Group of Governmental Experts to negotiate a legally-binding instrument to ensure meaningful human control over critical functions in lethal autonomous weapon systems"[16]. The primary contention underlying this and related proposals is often framed in the following terms: the exercise of control by machines — without sufficient capacity for human intervention, oversight, control, judgment, or some combination thereof — over decisions or conduct pertaining to the use of force in armed conflict is illegitimate and (should be) unlawful.

At the other end of the spectrum, certain States have expressed the position that existing IHL is sufficient in this area and that emerging technologies do not present insuperable barriers to compliance with IHL and the application of responsibility. Furthermore, in the view of at least some of these States, including the United States of America, the employment of weapons with an automatic or autonomous attribute may yield purported increases in capabilities to (among other things) enhance distinction between civilians and military objectives through greater precision and accuracy. According to that position, the employment of such technologies may result in greater protection of civilians, increased compliance with IHL, and heightened realization of some of the (other) humanitarian aims underlying IHL[17].

---

[13] CCW 2019. P. 10.

[14] Submission by Argentina to the GGE "Questionnaire on the Legal Review Mechanisms of New Weapons, Means and Methods of Warfare". March 29, 2019. URL: https://undocs.org/en/CCW/GGE.1/2019/WP.6 (accessed 29.04.2020); Submission by Australia to the GGE "The Australian Article 36 Review Process". August 30, 2018. URL: https://undocs.org/en/CCW/GGE.2/2018/WP.6 (accessed 29.04.2020); Submission by the United States to the GGE "Autonomy in Weapon System". November 10, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.6 (accessed 29.04.2020); Submission by the Netherlands and Switzerland to the GGE, *Weapons Review Mechanisms*. November 7, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.5 (accessed 29.04.2020). See also: Goussac N. Safety net or tangled web: Legal reviews of AI in weapons and war-fighting. – *ICRC Humanitarian Law and Policy Blog.* April 18, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/ (accessed 29.04.2020); Lewis D.A. Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider. – *ICRC Humanitarian Law and Policy Blog*. March 21, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/ (accessed 29.04.2020).

[15] Campaign to Stop Killer Robots "Country Views on Killer Robots". April 13, 2018. URL: https://www.stopkillerrobots.org/wp-content/uploads/2018/04/KRC_CountryViews_13Apr2018.pdf (accessed 29.04.2020); Statement by the Campaign to Stop Killer Robots to the GGE. March 27, 2019. URL:https://www.unog.ch/80256EDD006B8954/(httpAssets)/6FDBADBE15A26515C12583CB003E9791/$file/KRC_StmtCCW_27Mar2019_TODELIVER.pdf (accessed 29.04.2020).

[16] Submission by Austria, Brazil, and Chile to the GGE "Proposal for a Mandate to Negotiate a Legally-binding Instrument that addresses the Legal, Humanitarian and Ethical Concerns posed by Emerging Technologies in the Area of Lethal Autonomous Weapons Systems (LAWS)". August 29, 2018. URL: https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/documents/29August_Proposal_Mandate_ABC.pdf (accessed 29.04.2020).

[17] Submission by the United States to the GGE "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems". March 28, 2019. URL: https://undocs.org/en/CCW/GGE.1/2019/WP.5 (accessed 29.04.2020); Submission by the Russian Federation to the GGE, "Potential opportunities and limitations of military uses of lethal autonomous weapons systems". March 8, 2019. URL: https://www.unog.ch/80256EDD006B8954/(httpAssets)/489AAB0F44289865C12583BB0063B977/$file/GGE+LAWS+2019_Working+Paper+Russian+Federation_E.pdf (accessed 29.04.2020); Submission by the United States to the GGE "Humanitarian benefits of emerging technologies in the area of lethal autonomous weapon systems". April 3, 2018. URL: https://undocs.org/en/CCW/GGE.1/2018/WP.4 (accessed 29.04.2020); Submission by the United States to the GGE "Autonomy in Weapon System". November 10, 2017. URL: https://undocs.org/en/CCW/GGE.1/2017/WP.6 (accessed 29.04.2020).

3. A proposed analytical framework aimed at identifying general concepts and specific attributes necessary to apply international legal responsibility

Taking account of the rapid pace of technological developments and the current normative impasse in the GGE among certain States, in this section I propose an analytical framework through which to evaluate possible employments of AI-related technologies in armed conflict. In short, I submit that international actors ought to identify the general concepts and specific attributes necessary to apply international legal responsibility to States, international organizations, and individuals in this area. The identification of those general concepts and specific attributes across the diverse range of possibly implicated thematic areas — from weapons to detention to humanitarian services to warships — may provide a useful approach through which States and other international actors can seek to respect the law, guide behavior, and pursue accountability.

The proposed responsibility-focused framework might be especially relevant in respect of *a priori* assessments concerning the possible development, procurement, testing, fielding, or employment of particular AI-related technologies in armed conflict. By evaluating those technologies against the identified general concepts and specific attributes, States and other international actors may determine which of the technologies are capable of being employed in a manner that ensures the possibility to apply responsibility. In doing so, those actors may discern the necessary and sufficient conditions under which responsibility may be applied in respect of those technologies. Through this framework, international actors can also ascertain which employments of AI-related technologies are not capable of being used in a manner that permits the application of responsibility. By doing so, those actors may identify which employments may arguably be impermissible on that basis. For example, this analytical process can help enumerate what forms of predictability and reliability it may arguably be necessary to ensure — as well as what specific personal, temporal, geographical, and material constraints it may be necessary to impose — so that a particular employment of AI-related technologies in armed conflict respects the law[18]. Furthermore, by sharing their understandings — in the GGE or other fora — of the identified general concepts and specific attributes necessary to apply responsibility in this area, international actors can help generate areas of greater normative consensus and promote legal stability.

At least two objections to an approach that focuses solely or at least primarily on responsibility may be anticipated. First, because it focuses on the content of existing obligations, a responsibility-focused framework may arguably not reflect calls to elaborate new substantive primary obligations. That objection presumes, however, that those calls already provide sufficient detail as to what precisely such a new obligation would entail. Whether or not those calls do provide such detail is a matter of dispute. The process of thinking through what would constitute a comprehensive list of general concepts and specific attributes could yield a greater understanding of what is already legally required and, indeed, could do so at a level of granularity that is, by and large, currently lacking in international discourse. In turn, that understanding could help inform multilateral debate as to what elements of the existing legal framework are or are not necessary or sufficient to apply responsibility. It could also help inform the development of national measures aimed at respecting the law. In these ways, the responsibility-focused approach proposed here may be of value to all States, irrespective of their position with respect to the current normative impasse at the GGE.

A second possible objection relates to the notion that international law currently does not address the responsibility of non-State parties to armed conflict as comprehensively or systematically as it addressees the responsibility of States, international organizations, or individuals. Be that as it may, according to the principle of equality of belligerents, IHL applies equally to all parties to an armed conflict, including to non-State parties in non-international armed conflicts [Sassòli 2019:585–587]. In light of that principle, a comprehensive list of general concepts and specific attributes necessary to apply responsibility to States, international organizations, and individuals may be considered to also apply, mutatis mutandis, in respect of a non-State party to an armed conflict.

It bears emphasis that the following does not represent a comprehensive list of all of the possibly relevant general concepts and specific attributes necessary to apply responsibility in this area. Due to time limitations and space restrictions, I outline only

[18] Submission by the International Committee of the Red Cross to the GGE "Autonomy, artificial intelligence and robotics: Technical aspects of human control". August 20, 2019. URL: https://www.icrc.org/en/download/file/102852/autonomy_artificial_intelligence_and_robotics.pdf (accessed 29.04.2020).

a handful of examples, with a focus on the responsibility of States and individuals. I therefore do not address the responsibility of international organizations or other actors that may be considered to have international legal personality.

a. Responsibility of States

The underlying concepts of responsibility of States are general in character. Those concepts — namely, attribution, breach, excuses, and consequences — are assumed and apply unless excluded, for example through variances grounded in individual treaties or rules[19].

Regarding the general State-responsibility concept of *attribution*, it is axiomatic that, to apply responsibility, State conduct must be attributable to each relevant State. Accordingly, an employment of AI-related technologies in armed conflict by a State must be susceptible of being ascribed to that State. This aspect concerns the specific element of attributability. As an example, in situations of partnered warfare where two or more States are involved jointly in the employment of AI-related technologies to assist in the identification and prioritization of targets in an attack, the possibility to attribute that conduct to each relevant State may arguably need to be ensured. (Where an international organization is involved in partnered warfare, the rules of attribution concerning international organizations would (also) be implicated.)

Regarding the general State-responsibility concept of *breach*, an employment of AI-related technologies in armed conflict may implicate hundreds of primary obligations originating in IHL and other fields of law applicable in respect of armed conflict. Consider two of the several possible examples that may be drawn from IHL provisions.

First, under Article 48 of Additional Protocol I of 1977 and its customary-law counterpart, the parties shall at all times distinguish between the civilian population and combatants and between civilian objects and military objectives and accordingly shall direct their operations only against military objectives[20]. Extrapolating from that provision, the general State-responsibility concept of breach may arguably entail an obligation to ensure — with respect to military operations involving AI-related technologies — the specific attributes of: (i) discernibility of the civilian population; (ii) discernibility of combatants; (iii) discernibility of civilian objects; (iv) discernibility of military objectives; (v) distinguishability of (i) from (ii) and of (iii) from (iv); and (vi) capability to direct military operations only against (iv). An employment in a military operation of AI-related technologies lacking one (or more) of those specific attributes ok preclude the application of responsibility and may arguably be impermissible on that ground. Thus, for instance, an employment by a State of machine-learning techniques to assist in the collection or evaluation of intelligence pertaining to military operations may arguably need to entail those specific attributes. Whether — and, if so, the extent to which — AI-related technologies may or may not be relied upon to (help) make the evaluative decisions and normative judgments involved in specific attributes (i) through (v) is a key area that warrants attention.

Second, pursuant to Article 57, paragraph 2, lit. *b* of Additional Protocol I of 1977, "an attack shall be cancelled or suspended if it becomes apparent that the objective is not a military one or is subject to special protection or that the attack 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"[21]. Extrapolating with respect to this provision, the general State-responsibility concept of breach may arguably entail an obligation for the party to ensure — with respect to precautionary measures concerning an attack involving AI-related technologies — the specific attributes, where relevant, of: (i) cancellability; (ii) suspensibility; (iii) discernibility of (non-)military-objective status; (iv) discernibility of (non-)special-protection status; (v) discernibility of incidental loss of civilian life that may be caused; (vi) discernibility of injury to civilians that may be caused; (vii) discernibility of damage to civilian objects that may be caused; (viii) discernibility of a combination of incidental loss of civilian life, injury to civilians, and damage to civilian objects; (ix) discernibility of the concrete and direct military advantage anticipated; and (x) capability to evaluate whether (v), (vi), (vii), or (viii) would be excessive in relation to (ix). An employment by a State of AI-related technologies in an attack lacking one

[19] Crawford J.R. State Responsibility. – *Max Planck Encyclopedia of Public International Law.* 2006. URL: https://opil.ouplaw.com/view/10.1093/law:epil/9780199231690/law-9780199231690-e1093 (accessed 29.04.2020).

[20] Protocol additional to the Geneva Conventions of 12 August 1949, and relating to the protection of victims of international armed conflicts (Protocol I). Adopted June 8, 1977, entry into force December 7, 1978 (hereafter Additional Protocol I). URL: https://www.icrc.org/en/doc/assets/files/other/icrc_002_0321.pdf (accessed 29.04.2020).

[21] Article 57(2)(b) of the Additional Protocol I.

(or more) of those specific attributes would preclude the application of responsibility and may arguably be impermissible on that ground. Thus, for instance, the employment by a State of machine-learning techniques to assist in the identification of targets may arguably need to entail those specific attributes. Whether — and, if so, the extent to which — AI-related technologies may or may not be relied upon to (help) make the evaluative decisions and normative judgments involved in specific attributes (iii) through (x) is a key area that warrants attention regarding this set of specific attributes.

Regarding the general State-responsibility concept of legal *consequences*, attention may be drawn to cessation, which, according to the International Law Commission, is one of the two general consequences of an internationally wrongful act.[22] (Reparation is the other general consequence.) This general State-responsibility concept of cessation entails that a State responsible for an internationally wrongful act is under an obligation to cease that act if it is continuing, and, if circumstances so require, to offer appropriate assurances and guarantees of non-repetition.[23] The general State-responsibility concept of legal consequences thus concerns what could be characterized as the specific attributes of: (i) discernibility of an initial breach; and (ii) discernibility of wrongful-act continuation. For instance, an employment by a State of machine-learning techniques to assist in the identification and prioritization of targets in a series of attacks may arguably need to be undertaken in a manner that ensures the possibility to discern whether that employment constitutes an initial breach, and, if so, whether ongoing employments continue that breach. Given its nature, the general State-responsibility concept of consequences is arguably premised on ensuring the possibility of the exercise of external scrutiny. Accordingly, a possible guidepost to realize the specific attributes of (i) discernibility of initial breach and (ii) discernibility of wrongful-act continuation could be that an employment of AI-related technologies in armed conflict may arguably need to be capable of being subject to external juridical scrutiny.

Finally, regarding the general State-responsibility concepts of attribution, breach, and consequences, all conduct involving the employment of AI-related technologies in armed conflict concerns what might be characterized as the specific attribute of reconstructability. That attribute relates to the possibility to sufficiently piece together — before, during, and after employment — the inputs, functions, dependencies, and outputs of the computational components adopted, and by whom, in respect of such an employment[24]. A guidepost for the specific attribute of reconstructability in this area might be that such an employment is capable of being subject to juridical scrutiny, perhaps including by a judicial organ[25].

b. Responsibility of individuals under the international criminal law of war crimes

Certain general secondary rules of responsibility may be put forward, at least on a provisional basis, around responsibility of individuals under the international criminal law of war crimes. Those general individual-responsibility concepts might include attribution, prohibited conduct, mental elements, modes of responsibility, and penalties. It is arguable that those general concepts may be assumed and apply unless excluded, for example through the operation of a recognized and applicable excuse. (As a threshold matter, the imposition of individual responsibility under the international criminal law of war crimes assumes that a court may exercise jurisdiction in respect of a particular person, crime, territory, and period.) Some specific attributes may be identified by extrapolating from these general concepts and by drawing from particular textual provisions of the 1998 Rome Statute of the International Criminal Court (ICC Statute)[26]. For the purposes of the following analysis, it is assumed that the content of the excerpted provisions of the ICC Statute is applicable in respect of an employment of AI-related technologies in an armed conflict.

Regarding the general individual-responsibility concept of *attribution*, the ICC Statute lays down that the Court shall have jurisdiction over natural persons[27]. Accordingly, to ensure the possibility to

[22] International Law Commission: Draft articles on Responsibility of States for Internationally Wrongful Acts, with commentaries. – *Yearbook of the International Law Commission.* 2001. Vol. II. Part 2. P. 89.

[23] Ibid. P. 88.

[24] Lewis D.A. *Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider. – ICRC Humanitarian Law and Policy Blog.* March 21, 2019. URL: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/ (accessed 29.04.2020).

[25] Ibid.

[26] International Criminal Court: Rome Statute of the International Criminal Court. Adopted July 17, 1998, entry into force July 1, 2002 (hereafter ICC Statute). URL: https://www.icc-cpi.int/resource-library/documents/rs-eng.pdf (accessed 29.04.2020).

[27] Article 25(1) of the ICC Statute.

ascribe individual responsibility, an employment of AI-related technologies in armed conflict — at least one involved in conduct prohibited as a war crime — may arguably need to be susceptible of being attributed to one or more natural persons. This aspect thus concerns the specific element of attributability. Whether — and, if so, the extent to which — particular AI-related technologies may or may not be employed in a manner that facilitates the ascription of individual criminal responsibility is a key area that warrants greater attention.

Regarding the general individual-responsibility concept of *prohibited conduct*, the ICC Statute lists several forms of conduct (whether in the form of an act or an omission) that constitute war crimes so long as they are committed, as with all war crimes under the ICC Statute, as part of a plan or policy or as part of a large-scale commission of such crimes[28]. To ensure the possibility to apply the general individual-responsibility concept of prohibited conduct, an employment of AI-related technologies in armed conflict — at least one involved in conduct prohibited as a war crime — may arguably need to be susceptible of being assessed as to whether or not it satisfies the material elements of a possibly relevant form of prohibited conduct. There is a large array of forms of prohibited conduct under the international criminal law of war crimes. To enumerate which specific attributes would need to be ensured in respect of a particular employment of AI-related technologies in armed conflict, the material elements underlying each form of possibly relevant conduct prohibited as a war crime concerning that employment would need to be identified and transposed into specific attributes.

To give one of the numerous possible examples, under the ICC Statute the following conduct is characterized as a war crime when undertaken in an international armed conflict: attacking or bombarding, by whatever means, towns, villages, dwellings, or buildings which are undefended (in the sense of being open for unresisted occupation[29]) and which are not military objectives[30]. Extrapolating with respect to this provision, the general individual-responsibility concept of prohibited conduct may arguably entail an obligation for a relevant individual to ensure — with respect to an attack or bombardment that involves the employment of AI-related technologies and that possibly implicates at least one of the enumerated objects — the specific attributes of: (i) discernibility of towns; (ii) discernibility of villages; (iii) discernibility of dwellings; (iv) discernibility of buildings; (v) discernibility of whether an object at issue in (i)–(iv) is or is not open for unresisted occupation; (vi) discernibility of military-objective status; and (vii) capability to evaluate whether an object at issue in (i)–(iv) is or is not a military objective. An employment of AI-related technologies lacking one or more of these specific attributes, with respect to a relevant attack or bombardment, may arguably preclude the application of individual criminal responsibility and would be impermissible on that basis. For instance, consider the possibility that an individual contemplates employing machine-learning techniques to help identify targets in an attack against a building that is purportedly defended and that ostensibly constitutes a military objective. If those AI-related technologies lack the specific attributes (iv) through (vii), that employment would preclude the application of individual responsibility. Whether — and, if so, the extent to which — AI-related technologies may or may not be relied upon to (help) make the evaluative decisions and normative judgments involved in specific attributes (v) through (vii) is a key area that warrants attention regarding this set of specific attributes.

Concerning the general individual-responsibility concept of *mental elements*, under the ICC Statute a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court only if the material elements are committed with intent and knowledge, unless otherwise provided[31]. Under that provision, a person has intent where, in relation to conduct, that person means to engage in the conduct, and, in relation to a consequence, that person means to cause that consequence or is aware that it will occur in the ordinary course of events[32]. Also under that provision of the ICC Statute, knowledge means awareness that a circumstance exists or a consequence will occur in the ordinary course of events, and know and knowingly shall be construed accordingly[33]. Extrapolating from those provisions,

[28] Article 8(1) of the ICC Statute.

[29] Official Records of the Assembly of States Parties to the Rome Statute of the International Criminal Court. First session. September 3–10, 2002. P. 132. URL: https://documents-dds-ny.un.org/doc/UNDOC/GEN/N02/603/35/PDF/N0260335.pdf?OpenElement (accessed 29.04.2020).

[30] Article 8(2)(b)(v) of the ICC Statute.

[31] Article 30(1) of the ICC Statute.

[32] Article 30(2) of the ICC Statute.

[33] Article 30(3) of the ICC Statute.

at least two specific attributes arguably necessary to the application of individual responsibility with respect to mental elements concerning war crimes under the ICC Statute may be identified.

First, regarding the mental element of *intent*, a relevant employment of AI-related technologies in armed conflict — at least one involved in conduct prohibited as a war crime — may arguably need to facilitate a relevant natural person (or persons), in relation to conduct, to mean to engage in the conduct, or, in relation to a consequence, to mean to cause that consequence or to be aware that it will occur in the ordinary course of events. This aspect thus concerns the specific attributes of: (i) meaning to engage in certain conduct; (ii) meaning to cause a certain consequence; or (iii) being aware that a certain consequence will occur in the ordinary course of events. Whether — and, if so, the extent to which — certain AI-related technologies may or may not be employed in a manner that facilitates an exercise of those specific attributes concerning intent is another key area that warrants greater attention.

Second, with respect to the mental element of *knowledge*, an employment of AI-related technologies in armed conflict — at least one involved in conduct prohibited as a war crime — may arguably need to facilitate awareness either that a certain circumstance exists or that a certain consequence will occur in the ordinary course of events. This aspect thus concerns the specific aspects of: (i) being aware that the circumstance exists; or (ii) being aware that the consequence will occur in the ordinary course of events. Whether — and, if so, the extent to which — certain AI-related technologies may or may not be employed in a manner that facilitates an exercise of those specific attributes concerning knowledge is yet another key area that warrants greater attention.

Regarding the general individual-responsibility concept of *modes of responsibility*, under the ICC Statute a person shall be criminally responsible and liable for punishment for a crime within the jurisdiction of the Court under certain circumstances[34]. Those circumstances include where that person either commits such a crime (whether as an individual, jointly with another, or through another person, regardless of whether that other person is criminally responsible)[35] or where that person, for the purpose of facilitating the commission of such a crime, aids, abets, or otherwise assists in its commission or its attempted commission (including providing the means for its commission)[36]. Other modes of responsibility listed in the ICC Statute concern ordering, soliciting, or inducing the commission of a crime[37] or contributing to the commission or attempted commission of a crime in any other way by a group of persons acting with a common purpose[38]. Finally, the ICC Statute provides a basis for the Court to exercise jurisdiction on other grounds of criminal responsibility, namely certain forms of military-commander responsibility and (other forms of) superior responsibility[39].

To ensure the possibility to apply individual responsibility with respect to war crimes under the ICC Statute, an employment of AI-related technologies — at least where involved in conduct prohibited as war crimes — may arguably need to be susceptible of being assessed with respect to possibly relevant modes of responsibility. For example, it may be extrapolated, concerning situations involving possible facilitation of a war crime under the ICC Statute, that an employment of AI-related technologies may arguably need to permit the ascertainment of the following specific attributes: (i) discernibility of aiding; (ii) discernibility of abetting; and (iii) discernibility of otherwise assisting in the commission or attempted commission of a possibly relevant war crime.

Regarding the general individual-responsibility concept of *penalties*, the ICC Statute provides that the Court may impose imprisonment as well as order a fine or a forfeiture of proceeds, property, and assets derived directly or indirectly from the crime (or both)[40]. To ensure the possibility to apply individual responsibility in respect of war crimes under the ICC Statute, an employment of AI-related technologies — at least one involved in conduct prohibited as war crimes — may arguably need to be susceptible of the imposition of these penalties. This aspect thus concerns the specific attributes of: (i) possibility for a natural person to be imprisoned; (ii) possibility for a natural person to be fined; or (iii) (as relevant) forfeitability of proceeds, property, and assets.

[34] Article 25(3) of the ICC Statute.
[35] Article 25(3)(a) of the ICC Statute.
[36] Article 25(3)(c) of the ICC Statute.
[37] Article 25(3)(b) of the ICC Statute.
[38] Article 25(3)(d) of the ICC Statute.
[39] Article 28 of the ICC Statute.
[40] Article 77(1) of the ICC Statute.

4. Conclusion

One avenue to help identify and address legal concerns regarding the possible employment of AI-related technologies in armed conflict is for international actors to elaborate and share their understandings as to what would constitute a comprehensive list of the general concepts and specific attributes against which those employments should be assessed in order to ensure the applicability of responsibility. Doing so might strengthen the practical capacity of the law to regulate behavior in this area, enhance legal stability, and generate further areas of normative consensus among States.

References

1. Boulanin V., Verbruggen M. *Mapping the Development of Autonomy in Weapons Systems.* 2017. 147 p. URL: https://www.sipri.org/sites/default/files/2017-11/siprireport_mapping_the_development_of_autonomy_in_weapon_systems_1117_1.pdf (accessed 29.04.2020).
2. Ekelhof M., Persi Paoli G. *Swarm Robotics: Technical and Operational Overview of the Next Generation of Autonomous Systems.* Geneva: United Nations Institute for Disarmament Research. 2020. 8 p. URL: https://unidir.org/sites/default/files/2020-04/UNIDIR_Swarms_SinglePages_web.pdf (accessed 29.04.2020).
3. Ekelhof M.A.C. Lifting the Fog of Targeting: "Autonomous Weapons" and human control the lens of military targeting. – *Naval War College Review.* 2018. Vol. 73. No. 3. P. 61–94.
4. Kania E. *"AI Weapons" in China's Military Innovation*. 2020. 23 p. URL: https://www.brookings.edu/wp-content/uploads/2020/04/FP_20200427_ai_weapons_kania_v2.pdf (accessed 29.04.2020).
5. Lewis D.A., Modirzadeh N.K., Blum G. *War-Algorithm Accountability.* 2016. 245 p. URL: https://dash.harvard.edu/bitstream/handle/1/28265262/War-Algorithm-Accountability-August-2016.pdf?sequence=1&isAllowed=y (accessed 29.04.2020).
6. Nasu H., Letts D. The Legal Characterization of Lethal Autonomous Maritime Systems: Warship, Torpedo, or Naval Mine?. – *International Law Studies.* 2020. Vol. 96. P. 79–97.
7. Sassòli M. *International Humanitarian Law: Rules, Controversies, and Solutions to Problems Arising in Warfare.* Northampton, MA: Edward Elgar Publishing Limited. 2019. 656 p.
8. Sayler K. M. *Artificial Intelligence and National Security. Congressional Research Service. Report No. R45178.* November 21, 2019. 43 p. URL: https://fas.org/sgp/crs/natsec/R45178.pdf (accessed 29.04.2020).
9. Singer T. *Dehumanisierung der Kriegführung: Herausforderungen für das Völkerrecht und die Frage nach der Notwendigkeit menschlicher Kontrolle.* Berlin: Springer. 2018. 551 p.

About the Authors

Dustin Lewis,
LL.M., Research Director, Harvard Law School Program on International Law and Armed Conflict

175J, Massachusetts Avenue Langdell, Cambridge, MA, United States of America, 02138

dlewis@law.harvard.edu
ORCID: 0000-0003-3843-7665

Информация об авторах

Дастин Льюс,
LL.M, руководитель научно-исследовательских работ, Harvard Law School Program on International Law and Armed Conflict

02138, United States of America, Cambridge, MA, Massachusetts Avenue Langdell, 175J

dlewis@law.harvard.edu
ORCID: 0000-0003-3843-7665