INTERNATIONAL LAW STUDIES
Published Since 1895

Legal Reviews of War Algorithms

Tobias Vestner and Altea Rossi

97 INT'L L. STUD. 509 (2021)

Volume 97

2021

Published by the Stockton Center for International Law

ISSN 2375-2831

Legal Reviews of War Algorithms

Tobias Vestner[*] *and Altea Rossi*[**]

CONTENTS

* Tobias Vestner is Head of Security and Law Programme at the Geneva Center for Security Policy (GCSP), Fellow at Supreme Headquarters Allied Powers Europe, Honorary Senior Research Fellow at Exeter University, and Reserve Legal Adviser at the Swiss Armed Forces Staff. Email: t.vestner@gcsp.ch.

** Altea Rossi is Programme Officer at the Security and Law Programme at GCSP and Deputy Member to the Council of Europe Commission for Democracy through Law (Venice Commission). Email: a.rossi@gcsp.ch.

The authors thank Ashley S. Deeks, Michael W. Meier, Wen Zhou, Ricardo Chavarriaga, Benjamin Schumeg, Tarek Abulmagd, Adam Hilburn, Adam Hoxha, Newman Hsiao, Ryan Olsen, Katy Perez, Gagan Singh, and Carl Valianti for their valuable insights, as well as comments on a previous version of this article.

The thoughts and opinions expressed are those of the authors and not necessarily those of the U.S. government, the U.S. Department of the Navy, or the U.S. Naval War College.

I. INTRODUCTION

Remarkable developments in robotics over the last years have led to a new "summer" of artificial intelligence (AI).[1] Notably, machine learning and deep learning transform daily life. Humans increasingly rely on "external" intelligence without even realizing it.[2] The military has also recognized the significant potential of AI.[3] Security forces employ AI tools for information analysis and facial recognition, for instance. Yet, the interest goes further. Technologically advanced States, such as the United States, China, and Russia, have started to engage in an arms race regarding military applications of AI.[4] Major research and development projects, often involving partnerships between defense ministries, private companies, and academia, are currently

1. VINCENT BOULANIN ET AL., ARTIFICIAL INTELLIGENCE, STRATEGIC STABILITY AND NUCLEAR RISK 8 (2020), https://www.sipri.org/sites/default/files/2020-06/artificial_intelligence_strategic_stability_and_nuclear_risk.pdf [hereinafter ARTIFICIAL INTELLIGENCE, STRATEGIC STABILITY AND NUCLEAR RISK].

2. For example, smartphones and social media are powered by deep learning, a specific type of machine learning technique. Other civilian applications of AI may be automatic image recognition as well as AI applications for commercial purposes. For an overview of some civilian applications, see, e.g., *9 Applications of Machine Learning from Day-to-Day Life*, MEDIUM (July 31, 2017), https://medium.com/app-affairs/9-applications-of-machine-learning-from-day-to-day-life-112a47a429d0.

3. Paul D. Scharre, *The Opportunity and Challenge of Autonomous Systems*, *in* AUTONOMOUS SYSTEMS: ISSUES FOR DEFENSE POLICYMAKERS 3, 5–6 (Paul D. Scharre & Andrew P. Williams eds., 2015), https://www.act.nato.int/images/stories/media/capdev/capdev_02.pdf. *See also* Niklas Masuhr, *AI in Military Enabling Applications*, CSS ANALYSES IN SECURITY POLICY (Oct. 2019), https://css.ethz.ch/content/dam/ethz/special-interest/gess/cis/center-for-securities-studies/pdfs/CSSAnalyse251-EN.pdf. Despite a recent AI renaissance, it bears noting that AI applications have been used in the military domain since the 1960s, though in simpler forms. *See*, *e.g.*, STUART RUSSELL & PETER NORVIG, ARTIFICIAL INTELLIGENCE: A MODERN APPROACH (3d ed. 2014).

4. Carl B. Frey & Michael Osborne, *China Won't Win the Race for AI Dominance*, FOREIGN AFFAIRS (June 19, 2020), https://www.foreignaffairs.com/articles/united-states/2020-06-19/china-wont-win-race-ai-dominance; Tania Rabesandratana, *Europe Moves to Compete in Global AI Arms Race*, 360 SCIENCE 474, 474–75 (2018); Michael C. Horowitz, *The Algorithms of August*, FOREIGN POLICY (Sept. 12, 2018), https://foreignpolicy.com/2018/09/12/will-the-united-states-lose-the-artificial-intelligence-arms-race/; Tom Simonite, *For Superpowers, Artificial Intelligence Fuels New Global Arms Race*, WIRED (Sept. 8, 2017), https://www.wired.com/story/for-superpowers-artificial-intelligence-fuels-new-global-arms-race/; Edward M. Geist, *It's Already Too Late to Stop the AI Arms Race—We Must Manage it Instead*, 72 BULLETIN OF THE ATOMIC SCIENTISTS 318 (2016).

underway.[5] Given AI's significant advantages, there is a strong tendency for increased autonomy in security affairs. This includes an unbroken trend towards increased autonomy in relation to the military use of force against objects and persons.[6]

States and legal scholars have started to debate if and how military applications of AI might be compatible with existing international law, in particular international humanitarian law (IHL—here synonymously used as the "law of armed conflict" or the "law of war"). Mandated by the UN General Assembly, the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons has deliberated on this within the framework of the Convention on Certain Conventional Weapons (CCW) since 2016. In this context, several States have stressed the importance of legal reviews of weapons, means and methods of warfare according to Article 36 of Additional Protocol I to the 1949 Geneva Conventions[7] (API) and customary international law.[8] The legal review process, which assesses the legality of new weapons, would ensure that States do not employ AI-

5. *See*, *e.g.*, KELLEY M. SAYLER, CONG. RSCH. SERV., R45178, ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY 20–26 (Nov. 10, 2020).

6. The United States and the United Kingdom have stated that they will keep "human judgment" and "human control" over such systems: "[a]utonomous . . . weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force" and "operation of weapons systems will always be under human control," respectively. *See* U.S. Department of Defense, Directive 3000.09, Autonomy in Weapon Systems 13 (2012, incorporating Change 1, May 8, 2017) [hereinafter DoD Directive 3000.9]; UNITED KINGDOM MINISTRY OF DEFENSE, JDP 0-30.2, UNMANNED AIRCRAFT SYSTEMS 43, ¶ 4.18 (2017). Both terms ("human judgment" and "human control") are flexible notions that may lay themselves open to broad or even contradicting interpretations, however. *Cf.*, *e.g.*, Congressional Research Service, *Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems* (Dec. 1, 2019), https://fas.org/sgp/crs/natsec/IF11150.pdf; *Killer Robots: UK Government Policy on Fully Autonomous Weapons*, ARTICLE36 (Apr. 2013), http://www.article36.org/wp-content/uploads/2013/04/Policy_Paper1.pdf.

7. Protocol Additional to the Geneva Conventions of 12 August 1949, and Relating to the Protection of Victims of International Armed Conflicts art. 36, June 8, 1977, 1125 U.N.T.S. 3 [hereinafter API].

8. Several delegations have stressed explicitly the relevance of legal reviews of weapons in respect to lethal autonomous weapons in the context of the CCW, such as Brazil, Canada, the European Union, Greece, the Netherlands, the United Kingdom, and the United States. For specific references to such statements, see Eric T. Jensen, *The (Erroneous) Requirement for Human Judgment (and Error) in the Law of Armed Conflict*, 96 INTERNATIONAL LAW STUDIES 26, 40, 51 (2020).

powered systems that do not comply with IHL. Scholarly work has echoed the relevance of legal reviews regarding increasing autonomy related to weapon systems and have started to identify arising challenges. It has been noticed that new technology "can in some cases make the process of conducting an Article 36 review very difficult" and that this "might necessitate revising old legal concepts anew or pose new risks that may themselves require new methods of risk assessment."[9] In December 2019, the 33rd International Conference of the Red Cross and Red Crescent also stated that "for legal reviews to be effective, States that develop or acquire new weapon technologies need to navigate [their] complexities."[10]

This article answers this call for further reflection and digs deeper into the issue. It first provides an overview of emerging AI technology and its military applications, termed "war algorithms."[11] As such, this analysis applies to any type of operational use of AI that falls under the obligation to be legally reviewed, including its use in cyber operations, which inherently leads to a focus on the use of AI in relation to the conduct of hostilities. The article goes on to survey the debate among States in diplomatic fora and existing academic literature to outline the different perspectives on the legal review in the context of autonomous systems. It further analyzes in detail the obligation under IHL to conduct a legal review of weapons, means or methods of warfare, as well as the related State practice concerning such reviews. The article finds that while legal reviews are critical to prevent the deployment of weapons and systems that are non-compliant with existing international law, the existing practice is not fully adequate for reviewing the legality of AI-powered systems.

The article argues that States must adapt their legal reviews to the emerging AI technology. For AI systems that provide critical elements to human

9. VINCENT BOULANIN & MAAIKE VERBRUGGEN, ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES POSED BY EMERGING TECHNOLOGIES 6 (2017), https://www.sipri.org/sites/default/files/2017-12/article_36_report_1712.pdf [hereinafter ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES].

10. INTERNATIONAL COMMITTEE OF THE RED CROSS, INTERNATIONAL HUMANITARIAN LAW AND THE CHALLENGES OF CONTEMPORARY ARMED CONFLICTS: RECOMMITTING TO PROTECTION IN ARMED CONFLICT ON THE 70TH ANNIVERSARY OF THE GENEVA CONVENTIONS 29 (2019) [hereinafter ICRC, INTERNATIONAL HUMANITARIAN LAW AND THE CHALLENGES OF CONTEMPORARY ARMED CONFLICTS].

11. Dustin A. Lewis, Gabriella Blum & Naz K. Modirzadeh, *War-Algorithm Accountability*, at vii (Harvard Law School Program on International Law and Armed Conflict Research Briefing, 2016), http://blogs.harvard.edu/pilac/files/2016/09/War-Algorithm-Accountability-August-2016-compressed.pdf.

operators for targeting decisions and systems that autonomously take related decisions, legal reviews must assess the compliance with additional rules, in particular targeting law under IHL. Yet AI applications pose significant challenges regarding their predictability and explainability. This predictability problem is first and foremost an operational and technical challenge that can be addressed by the technical process of verification and validation, a process that generally precedes legal reviews. This article argues that for military systems that embed AI, as the law is translated into technical specifications, technical and legal assessments ultimately conflate into one. States thus need to conduct legal reviews as part of the technical validation and verification process. While this requires defining and assessing new parameters regarding predictability, among other consequences, the article suggests that emerging guidelines on the development and use of AI by States and industry can provide elements for the development of new guidance for the legal review of AI-driven systems. The article concludes that legal reviews become even more important for AI technology than for traditional weapons. With increased human reliance on AI, the legal review is the essential gatekeeper to its legal functioning.

II. MILITARY AND WEAPONIZED ARTIFICIAL INTELLIGENCE

AI can be described as a set of computational techniques that enables machines to solve complex and abstract problems that are naturally performed through human intelligence.[12] It can provide a system with "cognitive capabilities" to independently undertake functions such as observation, processing natural language, or learning.[13] Autonomy is therefore a constitutive

12. INTERNATIONAL PANEL ON THE REGULATION OF AUTONOMOUS WEAPONS (IPRAW), FOCUS ON COMPUTATIONAL METHODS IN THE CONTEXT OF LAWS (2017), https://www.ipraw.org/wp-content/uploads/2017/11/2017-11-10_iPRAW_Focus-On-Report-2.pdf; Vincent Boulanin, *Artificial Intelligence: A Primer*, *in* 1 THE IMPACT OF ARTIFICIAL INTELLIGENCE ON STRATEGIC STABILITY AND NUCLEAR RISK: EURO-ATLANTIC PERSPECTIVES 13, 13–14 (Vincent Boulanin ed., 2019), https://www.sipri.org/sites/default/files/2019-05/sipri1905-ai-strategic-stability-nuclear-risk.pdf [hereinafter IMPACT OF ARTIFICIAL INTELLIGENCE].

13. Boulanin, *Artificial Intelligence*, *supra* note 12, at 13; PAUL SCHARRE & MICHAEL C. HOROWITZ, ARTIFICIAL INTELLIGENCE: WHAT EVERY POLICYMAKER NEEDS TO KNOW 4 (2018), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS_AI_FINAL-v2.pdf?mtime=20180619100112&focal=none.

trait of AI.[14] Yet, the current systems' ability to know and act rests within predefined parameters that are set by human programmers. Current AI-driven systems perform only specified tasks to produce a desired outcome, such as visual perception, speech recognition, or decision-making. Accordingly, current AI techniques are called "narrow AI." In contrast, it is not yet possible for an AI system to pursue general purposes in general contexts as humans do, which is called "General Artificial Intelligence (AGI)."[15]

Current AI systems can perform both significantly worse and better than humans, depending on the task and context.[16] Regarding object recognition, for instance, deep-learning networks that learn how to recognize patterns of pixels making up images can misclassify what they "see" if just minimal changes occur in the pixel patterns. For this reason, they can be easily deceived through so-called "adversarial attacks." A team at Tesla, for instance, showed several ways to easily fool the autonomously driving cars' algorithms.[17] AI-based systems also still poorly perform for face-recognition, notably when it comes to certain ethnic groups,[18] and their reliance on software

14. Frank Sauer, *Military Applications of AI: Nuclear Risk Redux*, *in* IMPACT OF ARTIFICIAL INTELLIGENCE, *supra* note 12, at 84.

15. Ragnar Fjelland, *Why General Artificial Intelligence Will Not Be Realized*, HUMANITIES & SOCIAL SCIENCE COMMUNICATIONS (June 17, 2020), https://www.nature.com/articles/s41599-020-0494-4. *See also* Gideon Lewis-Kraus, *The Great A.I. Awakening*, NEW YORK TIMES MAGAZINE (Dec. 14, 2016), https://www.nytimes.com/2016/12/14/magazine/the-great-ai-awakening.html.

16. M.L. CUMMINGS, ARTIFICIAL INTELLIGENCE AND THE FUTURE OF WARFARE 8 (2017).

17. Furthermore, scientists from the Massachusetts Institute of Technology have shown the potential of "adversarial attacks." They demonstrated how reconnaissance systems based on deep learning can be easily led to misclassify objects, such as a turtle with a rifle. *See, e.g.*, Will Knight, *Military Artificial Intelligence Can Be Easily and Dangerously Fooled*, MIT TECHNOLOGY REVIEW (Oct. 21, 2019), https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/.

18. For example, AI has demonstrated poor performance in face-recognition of people with darker skin, which has recently led IBM, Amazon, and Microsoft to set self-imposed moratoriums on such technologies. *See* Karen Weise & Natasha Singer, *Amazon Pauses Police Use of Its Facial Recognition Software*, NEW YORK TIMES (June 10, 2020), https://www.nytimes.com/2020/06/10/technology/amazon-facial-recognition-backlash.html; Jay Greene, *Microsoft Won't Sell Police its Facial-Recognition Technology, Following Similar Moves by Amazon and IBM*, WASHINGTON POST (June 11, 2020), https://www.washingtonpost.com/technology/2020/06/11/microsoft-facial-recognition/.

makes them particularly vulnerable to cyber-attacks.[19] Yet, the advantages of AI for the conduct of military operations are undeniable for tasks where speediness of reaction is essential (such as cyber operations), or a vast amount of data needs to be managed. AI techniques particularly allow militaries to undertake a wide range of tasks with more accuracy and effectiveness than humans, notably surveillance operations through computer vision,[20] fighter pilot training, and search and rescue operations.[21]

Given these significant advantages, AI is increasingly used for military operations, especially in relation to targeting. Targeting is the engagement of an object or person with detrimental kinetic or non-kinetic effects. The U.S. Department of Defense (DoD) defines targeting as "the process of selecting and prioritizing targets and matching the appropriate response to them, considering operational requirements and capabilities," the purpose of which is to "integrate and synchronize fires into joint operations by utilizing available capabilities to generate a specific lethal or nonlethal effect on a target."[22] Applied to targeting functions, AI may serve as a tool to "empower unmanned systems to perform critical missions safely and with high degree of autonomy," as the U.S. Defense Advanced Research Projects Agency (DARPA) writes.[23] The replacement of human operators by AI systems in targeting cycles may further allow greater speed in observing, orienting, deciding, and acting, as well as increased accuracy for hitting a target, the ability

19. *See* Gheorghe Calopăreanu, *Aspects of Employing Artificial Intelligence in the Fighting Area*, 10 ANNALS: SERIES ON MILITARY SCIENCE, Issue No. 2, at 31, 36 (2018).

20. For example, the U.S. Project Maven attempted to create computer vision AI to analyze vast amount of surveillance footage. *See* David Herron, *Project Maven to Deploy Computer Algorithms to War Zone by Year's End*, TECH SPARX (July 20, 2017), https://techsparx.com/blog/2017/07/project-maven.html.

21. Robert W. Button, *Artificial Intelligence and the Military*, REAL CLEAR DEFENSE (Sept. 7, 2017), https://www.realcleardefense.com/articles/2017/09/07/artificial_intelligence_and_the_military_112240.html; Naveen Joshi, *4 Ways Global Defense Forces Use AI*, FORBES (Aug. 26, 2018), https://www.forbes.com/sites/cognitiveworld/2018/08/26/4-ways-the-global-defense-forces-are-using-ai/#37307f5503e4.

22. *See* U.S. Joint Chiefs of Staff, JP 3-60, Joint Targeting, at I-1, I-6 (2013), https://www.justsecurity.org/wp-content/uploads/2015/06/Joint_Chiefs-Joint_Targeting_20130131.pdf.

23. *AI Next Campaign*, DARPA, https://www.darpa.mil/work-with-us/ai-next-campaign (last visited Feb. 22, 2021).

to maintain unaltered performance over time,[24] and extended reach.[25] Such potential has been recently demonstrated during simulations of a dogfight between two F-16 fighter jets, for instance, in which the human pilot lost against an AI-powered algorithm.[26]

Autonomy in targeting is already a reality. Relatively simple rule-based control software undertakes targeting functions without human intervention. This is the case for air-defense systems or missiles with autonomous modes. The U.S. Air Force CQM-121A Pave Tiger and the YGCM-121B Seek Spinner, both unmanned aerial vehicles with anti-radar munitions, for instance, constitute such autonomous weapons, yet lack AI.[27] Distinguished from such systems that engage in the targeting decision themselves are AI tools that support the targeting cycle. AI-enabled intelligence, surveillance, and reconnaissance (ISR) systems provide prediction and identification, for instance. In this case, while the data provided by the algorithm is essential to reaching a targeting decision, the ultimate decision rests with the human operator.[28]

In diplomatic fora, notably the CCW, the debate has been almost exclusively focused on a more advanced type of AI application for targeting, so-called "Lethal Autonomous Weapons Systems" (LAWS). The United States defines these systems as those that, once activated, "can select and engage

24. Scharre, *supra* note 3, at 3. On this issue, in the context of the Group of Governmental Experts on Lethal Autonomous Weapons Systems meetings, Russia was of the view that LAWS "are capable of considerably reducing the negative consequences of the use of weapons related to operator's errors, mental and physiological state, as well as ethical, religious or moral stance in the IHL context." Russian Federation, Potential Opportunities and Limitation of Military Uses of Lethal Autonomous Weapons Systems 4, U.N. Doc. CCW/GGE.1/2019/WP.1 (Mar. 15, 2019), https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2019/gge/Documents/GGE.2-WP1.pdf.

25. GREG L. ZACHARIAS, AUTONOMOUS HORIZONS: THE WAY FORWARD 8 (2019), https://www.airuniversity.af.edu/Portals/10/AUPress/Books/b_0155_zacharias_autonomous_horizons.pdf.

26. Patrick Tucker, *An AI Just Beat a Human F-16 Pilot in a Dogfight—Again*, DEFENSE ONE (Aug. 20, 2020), https://www.defenseone.com/technology/2020/08/ai-just-beat-human-f-16-pilot-dogfight-again/167872/.

27. DEFENSE INNOVATION BOARD, U.S. DEPARTMENT OF DEFENSE, AI PRINCIPLES: RECOMMENDATIONS ON THE ETHICAL USE OF ARTIFICIAL INTELLIGENCE BY THE DEPARTMENT OF DEFENSE 12 (2019), https://media.defense.gov/2019/Oct/31/2002204459/-1/-1/0/DIB_AI_PRINCIPLES_SUPPORTING_DOCUMENT.PDF.

28. As to how AI applications can help armed forces in reaching targeting decisions, see, e.g., Ashley Deeks, *Coding the Law of Armed Conflict: First Steps*, *in* THE LAW OF ARMED CONFLICT IN 2040 (Matthew C. Waxman ed., forthcoming 2021) (manuscript at 4-5).

targets without further intervention by a human operator."[29] Disagreement exists on how to conceive the notion of autonomy.[30] Yet, it is generally agreed that AI has not reached the stage of enabling a system to operate fully autonomously during the entire targeting phase, namely making its own decisions as to whom to target and when to fire weapons.[31] It is generally agreed that such systems do not yet exist.[32] Some types of close-in weapons systems, such as the U.S. Phalanx, are able to detect and evaluate threats on their own, as well as track, engage, and destroy them.[33] Yet, such systems are employed in very limited contexts and remain under real-time human supervision.[34]

---

29. DoD Directive 3000.9, *supra* note 6, at 13–14. Similarly, the International Committee of the Red Cross defines LAWS as "[a]ny weapon system with autonomy in its critical functions. That is, a weapon system that can select (i.e. search for or detect, identify, track, select) and attack (i.e. use force against, neutralize, damage or destroy) targets without human intervention." International Committee of the Red Cross, *Views of the International Committee of the Red Cross (ICRC) on Autonomous Weapon System* (Apr. 11, 2016), https://www.unog.ch/80256EDD006B8954/(httpAssets)/B3834B2C62344053C1257F9400491826/%24file/2016_LAWS+MX_CountryPaper_ICRC.pdf. *See also* Christof Heyns, (Special Rapporteur on Extrajudicial, Summary or Arbitrary Executions), *Report*, ¶ 38, U.N. Doc. HRC/23/47 (Apr. 9, 2013), https://undocs.org/A/HRC/23/47. For an overview of LAWS definitions, *see* Michael C. Horowitz, *Why Words Matter: The Real World Consequences of Defining Autonomous Weapons Systems*, 30 TEMPLE INTERNATIONAL & COMPARATIVE LAW JOURNAL 85, 86-87 (2016).

30. Generally, autonomy refers to the ability of a machine to execute certain tasks without human input, using interactions of computer programming with the environment. ARTIFICIAL INTELLIGENCE, STRATEGIC STABILITY AND NUCLEAR RISK, *supra* note 1, at 13 (quoting Andrew Williams, *Defining Autonomy in Systems: Challenges and Solutions*, *in* AUTONOMOUS SYSTEMS: ISSUES FOR DEFENCE POLICYMAKERS 27 (Andrew Williams & Paul Scharre eds., 2015), https://www.act.nato.int/images/stories/media/capdev/capdev_02.pdf. On autonomy and weapons systems, *see generally* John O. Birkeland, *The Concept of Autonomy and the Changing Character of War*, 5 OSLO LAW REVIEW 73, 86 (2018); PAUL SCHARRE & MICHAEL C. HOROWITZ, AN INTRODUCTION TO AUTONOMY IN WEAPON SYSTEMS (2015). On autonomy in targeting, see *infra* Part V.

31. Michael C. Horowitz, *The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons*, 145 DAEDALUS 25, 27 (2016).

32. *See, e.g.,* Rebecca Crootof, *The Killer Robots Are Here: Legal and Policy Implications*, 36 CARDOZO LAW REVIEW 1837, 1863 (2015) (stating that "[t]here is a nearly universal agreement . . .that [LAWS] do not yet exist.").

33. Rajesh Uppal, *Close-in Weapons Systems (CIWS) to Provide Last Chance Defence till Replaced by Railguns and Lasers*, INTERNATIONAL DEFENCE, SECURITY & TECHNOLOGY (Aug. 2, 2019), https://idstch.com/military/navy/close-in-weapons-system-ciws-to-provide-last-chance-defence-till-replaced-by-railguns-and-lasers/.

34. SCHARRE & HOROWITZ, *supra* note 30, footnote 51 and accompanying text.

States' interest in employing AI for managing military operations, most notably targeting cycles, is likely to increase as technologies continue to advance. Advances in machine learning and deep learning have led ministries of defense to direct their research and development (R&D) at increasing the autonomy of weapon systems through AI. The United States, China, Russia, the United Kingdom, France, Israel, and South Korea are investing considerable resources in R&D projects to employ AI in autonomous targeting.[35] The United States established autonomy as one of the main pillars of its military development in its "Vision for Air Force Science and Technology 2010-2030."[36] DARPA aims to increase autonomy in military applications and create "machines [that] are more than just tools that execute human-programed rules or generalize from human-curated data set." In other words, these machines are intended rather as "colleagues than . . . tools."[37] China massively invests in new partnerships with private commercial entities and academic actors to "use advance commercial technology to serve the military."[38] China also rapidly progresses in automating weapons through AI, as illustrated by its recent development of cutting-edge drones.[39] Russia established close cooperation between the public and private sector,[40] and in 2012, it created an institution analogous to the U.S. DARPA to promote R&D in military technology.[41] More recently, Russia identified AI as a means

35. FRANK SLIJPER, ALICE BECK & DAAN KAYSER, STATE OF AI: ARTIFICIAL INTELLIGENCE, THE MILITARY, AND INCREASINGLY AUTONOMOUS WEAPONS (2019).

36. WERNER J.A. DAHM, REPORT ON TECHNOLOGY HORIZONS: A VISION FOR AIR FORCE SCIENCES AND TECHNOLOGY DURING 2010-2030, at iv (2010).

37. *AI Next Campaign*, *supra* note 23. It is worth noting, that the U.S. Army has also set up "a campaign of learning to aggressively pursue an Artificial Intelligence and machine learning-enabled battlefield management system," called "Project Convergence." The project is intended to transform the Army into "a Multi-Domain Force by 2035." *See Project Convergence*, ARMY FUTURES COMMAND, https://armyfuturescommand.com/convergence/ (last visited Feb. 22, 2021).

38. SLIJPER, BECK & KAYSER, *supra* note 35, at 13–15.

39. GREGORY C. ALLEN, UNDERSTANDING CHINA'S AI STRATEGY: CLUES TO CHINESE STRATEGIC THINKING ON ARTIFICIAL INTELLIGENCE AND NATIONAL SECURITY (2019), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS-Understanding-Chinas-AI-Strategy-Gregory-C.-Allen-FINAL-2.15.19.pdf?mtime=20190215104041&focal=none.

40. Samuel Bendett, *Here's How the Russian Military Is Organizing to Develop AI*, DEFENSE ONE (July 20, 2018), https://www.defenseone.com/ideas/2018/07/russian-militarys-ai-development-roadmap/149900/.

41. *See The Foundation for Advanced Research Projects*, https://fpi.gov.ru/ (last visited Feb. 22, 2021).

to control autonomous military systems.[42] It also reportedly developed an autonomous drone that could "take off, accomplish its mission, and land without human interference."[43]

III. THE DIPLOMATIC AND LEGAL DEBATE

Multilateral diplomacy and legal scholarship have started to address the rising challenges of increased autonomy related to targeting. Ethical and legal concerns over autonomous systems that make decisions over life and death of humans have led to diplomatic initiatives and deliberations within the CCW. Since 2014 the CCW has started to dedicate working sessions on the topic, ultimately establishing a Group of Governmental Experts on LAWS that meets biannually. The Group of Governmental Experts' purpose is to delve into critical questions related to the possible development and deployment of LAWS. To this end, States submit working papers, share their practice, and present proposals. While the Group of Governmental Experts summarizes its discussions and proposes the way forward every year in a report, the debate on LAWS continues at the First Committee on Disarmament and International Security of the UN General Assembly.[44]

The deliberations among States initially centered around whether to ban autonomous weapons. Only a few States took a clear stance that such a ban would be necessary. As of October 2019, within the First Committee, thirty States favored a preemptive ban.[45] Notably, China expressed the desire to

42. Samuel Bendett, *In AI, Russia Is Hustling to Catch Up*, DEFENSE ONE (Apr. 4, 2018), https://www.defenseone.com/ideas/2018/04/russia-races-forward-ai-development/147178/.

43. SLIJPER, BECK & KAYSER, *supra* note 35, at 17 (quoting Kyle Mizokami, *This is Russia's First Autonomous Strike Drone*, POPULAR MECHANICS (Jan. 25, 2019), https://www.popularmechanics.com/military/aviation/a26027921/russia-autonomous-strike-drone-okhotnik/).

44. *See*, *e.g.*, Meetings Coverage, General Assembly, First Committee Weighs Potential Risks of New Technologies as Members Exchange Views on How to Control Lethal Autonomous Weapons, Cyberattacks, U.N. Meetings Coverage GA/DIS/3611 (Oct. 26, 2018), https://www.un.org/press/en/2018/gadis3611.doc.htm.

45. Algeria, Argentina, Austria, Bolivia, Brazil, Chile, China, Colombia, Costa Rica, Cuba, Djibouti, Ecuador, Egypt, El Salvador, Ghana, Guatemala, Holy See, Iraq, Jordan, Mexico, Morocco, Namibia, Nicaragua, Pakistan, Panama, Peru, State of Palestine, Uganda, Venezuela, and Zimbabwe. *Country Views on Killer Robots*, CAMPAIGN TO STOP KILLER ROBOTS (Oct. 25, 2019), https://www.stopkillerrobots.org/wp-content/uploads/2019/10/KRC_CountryViews_25Oct2019rev.pdf.

negotiate and conclude a protocol to ban the use of fully autonomous LAWS without banning their development.[46] The adoption of a legally binding ban of LAWS is not expected in the near future however, as Australia, France, Israel, the Republic of Korea, Russia, Turkey, the United States, and the United Kingdom have opposed the negotiation of such a treaty.[47]

The debate on LAWS has sparked critical reflections on whether the current international legal framework, in particular IHL, is suitable for regulating these new technologies and their use in wartime. Some States argue that IHL constitutes a comprehensive and sufficient framework and that further norms or regulations setting boundaries for the use of military AI systems are unnecessary.[48] Russia and the United States are the main supporters of such a position, which has been termed an "apply and comply" or "wait-and-see" approach.[49] Other States argued the contrary, namely that this "can hardly solve, in a fundamental way, the concerns."[50] Others demonstrated their receptiveness for additional principles, guidelines, or codes of conduct for operating such systems.[51] The proposition to adopt a respective code of

46. Elsa Kania, *China's Strategic Ambiguity and Shifting Approach to Lethal Autonomous Weapon Systems*, LAWFARE (Apr. 17, 2018), https://www.lawfareblog.com/chinas-strategic-ambiguity-and-shifting-approach-lethal-autonomous-weapons-systems.

47. *Country Views on Killer Robots*, *supra* note 45. Some of them, however, at least supported a politically binding instrument, such as France. Ray Acheson, *New Law Needed Now*, REACHING CRITICAL WILL 1, 2 (Aug. 30, 2018), https://reachingcriticalwill.org/images/documents/Disarmament-fora/ccw/2018/gge/reports/CCWR6.9.pdf.

48. *Country Views on Killer Robots*, *supra* note 45.

49. Rebecca Crootof, *Regulating New Weapons Technology*, *in* THE IMPACT OF EMERGING TECHNOLOGIES ON THE LAW OF ARMED CONFLICT 4, 21 (Eric Talbot Jensen & Ronald T. P. Alcala eds., 2019) [hereinafter IMPACT OF EMERGING TECHNOLOGIES]. For a summary of the positions upheld during LAWS expert meetings, see Group of Government Experts of the High Contracting Parties to the CCW, Report of the 2018 Group of Governmental Experts on Lethal Autonomous Weapons Systems, ¶ 28, U.N. Doc. CCW/GGE.1/2018/3 (Oct. 23, 2018), https://undocs.org/en/CCW/GGE.1/2018/3.

50. China, Position Paper, ¶ 4, U.N. Doc. CCW/GGE.1/2018/WP.7 (Apr. 11, 2018), https://undocs.org/en/CCW/GGE.1/2018/WP.7.

51. While States of the Non-Aligned Movement and of the African Group, plus Austria, Brazil and Mexico favor the negotiation of a legally binding instrument, States which support or at least are "open to discuss further" a politically binding instrument are Australia, Belgium, Germany, Finland, France, Ireland, Italy, Norway, Poland, Spain, Sri Lanka, Sweden, and Switzerland. They rely on the argument that a legally binding instrument would be "premature." *See* Acheson, *supra* note 47, at 2.

conduct has not led to any concrete result yet.[52] In this context, a group of experts independently met in 2019 to discuss practical, legal, ethical, and operational considerations presented by LAWS with the goal of producing a list of "Guiding Principles for the Development and Use of LAWS" as a potential starting point towards good international practice.[53] At the same time, States have started adopting guidance for the development and use of AI for security purposes.[54]

Both States that argue that existing IHL is a sufficient legal framework for such new technologies and States that argue that additional rules are necessary have stressed the importance of legal reviews of weapons, means and methods of warfare to ensure new technologies' compliance with IHL. Indeed, as will be discussed below, legal reviews serve as a bulwark against fielding systems that cannot comply with IHL. The United States, among others, has emphasized the relevance of legal reviews. It has proposed that States develop "best practices" to conduct legal reviews of autonomous weapon systems.[55] Recently, Argentina proposed to produce a compendium of good practices of legal reviews, focusing on the acquisition phase.[56] The

52. Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons System, Report of the 2019 Session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, ¶ 5, U.N. Doc. CCW/GGE.1/2019/3 (Sept. 25, 2019), https://undocs.org/en/CCW/GGE.1/2019/3 [hereinafter Report of the 2019 Session].

53. The Canberra Working Group, *Guiding Principles for the Development and Use of LAWS: Version 1.0*, E-INTERNATIONAL RELATIONS (Apr. 15, 2020), https://www.e-ir.info/2020/04/15/guiding-principles-for-the-development-and-use-of-laws-version-1-0/.

54. For example, the United States adopted ethical principles on AI in October 2019 (on such guidance, see *infra* Part VIII). France also adopted its own AI strategy for the defense sector, to date the only one concerning the security sector to have been publicly released. *See* VINCENT BOULANIN ET AL., RESPONSIBLE MILITARY USE OF ARTIFICIAL INTELLIGENCE: CAN THE EUROPEAN UNION LEAD THE WAY IN DEVELOPING BEST PRACTICE? 8 (2020), https://sipri.org/sites/default/files/2020-11/responsible_military_use_of_artificial_intelligence.pdf (quoting FRENCH MINISTRY OF THE ARMED FORCES, L'INTELLIGENCE ARTIFICIELLE AU SERVICE DE LA DÉFENSE: RAPPORT DE LA TASK FORCE IA [ARTIFICIAL INTELLIGENCE AT THE SERVICE OF DEFENSE: REPORT OF THE AI TASK FORCE] (2019)).

55. Michael W. Meier, *Lethal Autonomous Weapons Systems (LAWS): Conducting a Comprehensive Legal Review*, 30 TEMPLE INTERNATIONAL & COMPARATIVE LAW JOURNAL 119, 123 (2016).

56. Argentina, Questionnaire on the Legal Review Mechanisms of New Weapons, Means and Methods of Warfare, ¶¶ 1–2, U.N. Doc. CCW/GGE.1/2019/WP.6 (Mar. 29,

Group of Governmental Experts ultimately concluded at its second session in 2019 that good practices in the conduct of the legal reviews of LAWS is one of the issues "that may benefit from additional clarification."[57] In December 2019, the 33rd International Conference of the Red Cross and Red Crescent stressed again the role and application of legal reviews to emerging technologies.[58]

While the diplomatic debate has been rather uncritical towards the application of legal reviews to AI systems, few legal practitioners and scholars have studied the issue in depth. Initial works have applied existing State practices to autonomous systems, thereby asserting that IHL has an appropriate mechanism to ensure compliance by new technologies.[59] Boulanin and Verbruggen, researchers at the Stockholm International Peace Research Institute, have built on this to map existing practices and identify challenges regarding the legal reviews of emerging technologies. They concluded that for weapons that can operate autonomously through AI applications, the existing practice for conducting legal reviews would not be sufficient. AI systems would need to be checked differently than traditional weapons to ensure their compliance with international law.[60] Other commentators in a blog series of the International Committee of the Red Cross (ICRC) specifically

---

2019), https://unog.ch/80256EDD006B8954/(httpAssets)/52C72D09DCA60B8BC125841E003579D8/$file/CCW_GGE.1_2019_WP.6.pdf.

57. Report of the 2019 Session, *supra* note 52, ¶ 18.c.

58. *See* ICRC, INTERNATIONAL HUMANITARIAN LAW AND THE CHALLENGES OF CONTEMPORARY ARMED CONFLICTS, *supra* note 10, at 34–35. Already in 2003, at the 28th International Conference of the Red Cross and Red Crescent, it was reaffirmed by consensus the need to ensure "the legality of new weapons under international law . . . in light of the rapid developments of weapons technology and in order to protect civilians from the indiscriminate effects of weapons and combatants from unnecessary suffering and prohibited weapons." *See* INTERNATIONAL COMMITTEE OF THE RED CROSS & INTERNATIONAL FEDERATION OF RED CROSS AND RED CRESCENT SOCIETIES, 28TH INTERNATIONAL CONFERENCE OF THE RED CROSS AND RED CRESCENT 20, https://www.icrc.org/en/doc/assets/files/other/icrc_002_1103.pdf (last visited Feb. 22, 2021).

59. Meier, *supra* note 55. William H. Boothby, *Highly Automated and Autonomous Technologies*, *in* NEW TECHNOLOGIES AND THE LAW IN WAR AND PEACE 137 (William H. Boothby ed., 2018) [hereinafter NEW TECHNOLOGIES].

60. VINCENT BOULANIN & MAAIKE VERBRUGGEN, STOCKHOLM INTERNATIONAL PEACE RESEARCH INSTITUTE COMPENDIUM ON ARTICLE 36 REVIEWS 16 (2017), https://www.sipri.org/sites/default/files/2017-12/sipri_bp_1712_article_36_compendium_2017.pdf [hereinafter COMPENDIUM ON ARTICLE 36 REVIEWS].

dedicated to LAWS underlined the value of legal reviews.[61] This work echoes the importance given to legal reviews by States in the context of the CCW. It also sets the ground for assessing in detail how to conduct legal reviews of military AI systems in practice and examines intrinsic challenges related to this emerging technology.

IV. THE LEGAL REVIEW OF WEAPONS, MEANS OR METHODS OF WARFARE

Legal reviews arise out of IHL's requirement for States to assess if new weapons, means or methods of warfare are prohibited in some or all circumstances by international law.[62] The purpose of legal reviews is to avert the deployment of weapons that have been banned or restricted by specific international legal rules or are not capable of complying with primary rules regulating the conduct of hostilities. Hence, legal reviews are national tools to prevent violations of international law that might occur with the introduction and use of new weapons.[63]

States parties to API are bound by the obligation to conduct legal reviews under Article 36. This obligation builds on other IHL provisions, notably the preamble of the 1868 Saint Petersburg Declaration relating to explosive

61. Netta Goussac, *Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-fighting*, HUMANITARIAN LAW & POLICY (Apr. 18, 2019), https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting/; Dustin A. Lewis, *Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider*, HUMANITARIAN LAW & POLICY (Mar. 21, 2019), https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider/. For a further analysis on autonomous weapons and legal reviews, see Nikolas Stürchler & Michael Siegrist*, A "Compliance-Based" Approach to Autonomous Weapon Systems,* EJIL:TALK! (Dec. 1, 2017), https://www.ejiltalk.org/a-compliance-based-approach-to-autonomous-weapon-systems/.

62. *See* COMMENTARY ON THE ADDITIONAL PROTOCOLS OF 8 JUNE 1977 TO THE GENEVA CONVENTIONS OF 12 AUGUST 1949, ¶ 1469 (Yves Sandoz, Christophe Swinarski & Bruno Zimmermann eds., 1987).

63. KATHLEEN LAWAND, INTERNATIONAL COMMITTEE OF THE RED CROSS, A GUIDE TO THE LEGAL REVIEW OF WEAPONS, MEANS AND METHODS OF WARFARE: MEASURES TO IMPLEMENT ARTICLE 36 OF ADDITIONAL PROTOCOL I OF 1977, at 4 (2006) [hereinafter ICRC GUIDE].

projectiles[64] and Article 1 common to the 1949 Geneva Conventions.[65] In addition, the Hague Convention IV Regulations and API affirm that the right of belligerents to choose and adopt means and methods of warfare "is not unlimited."[66] This implies a duty of care to assess the legality of weapons being developed and intended to be used.[67] The obligation to conduct legal reviews in API is formulated in general and vague terms. Article 36 states that:

> In the study, development, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.[68]

---

64. Declaration Renouncing the Use, in Time of War, of Explosive Projectiles Under 400 Grammes Weight, Nov. 29/Dec. 11, 1868, 138 Consol. T.S. 297, 18 MARTENS NOUVEAU RECUEIL (ser. 1) 474 [hereinafter 1868 St. Petersburg Declaration].

65. William H. Boothby, *Regulating New Weapon Technologies*, *in* NEW TECHNOLOGIES, *supra* note 59, at 16, 17. *See also* ICRC GUIDE, *supra* note 63, at 4. According to the 1868 St. Petersburg Declaration,

> the Contracting or Acceding Parties reserve to themselves to come hereafter to an understanding whenever a precise proposition shall be drawn up in view of future improvements which science may effect in the armament of troops, in order to maintain the principles which they have established, and to conciliate the necessities of war with the laws of humanity.

1868 St. Petersburg Declaration, *supra* note 64. Article 1 common to the 1949 Geneva Conventions provides, "[t]he High Contracting Parties undertake to respect and to ensure respect for the present Convention in all circumstances." *See, e.g.*, Convention (I) for the Amelioration of the Condition of the Wounded and Sick in the Armed Forces in the Field art. 1, Aug. 12, 1949, 6 U.S.T. 3114, 75 U.N.T.S. 31. On Common Article 1, see also TALLINN MANUAL ON THE INTERNATIONAL LAW APPLICABLE TO CYBER OPERATIONS 153 (Michael N. Schmitt ed., 2013).

66. Regulations Respecting the Laws and Customs of War on Land, annexed to Convention No. IV Respecting the Laws and Customs of War on Land art. 22, Oct. 18, 1907, 36 Stat. 2227, T.S. No. 539 [hereinafter Hague Regulations]; API, *supra* note 7, art. 35(1).

67. P.J. Blount, *The Preoperational Legal Review of Cyber Capabilities: Ensuring the Legality of Cyber Weapons*, 39 NORTHERN KENTUCKY LAW REVIEW 214 (2012).

68. API, *supra* note 7, art. 36.

Whether Article 36 has attained customary nature remains debated.[69] Yet, States not party to API, such as the United States and Israel, have institutionalized legal reviews. The U.S. practice even preceded the adoption of API. In any case, a narrower obligation exists under customary law imposing a responsibility on States to ensure at least that means of warfare to be acquired or used are compliant with IHL.[70] Article 36 and the related customary norm are silent regarding the procedures to be followed, the consequences of the review's findings, and which actor within a State must conduct legal reviews.[71] Accordingly, how to conduct legal reviews is mostly defined by national regulations, policy, and practice.[72] The United States and the United Kingdom are leading in this regard. They, as well as Australia, Norway, and Sweden, are among the few States that have shared their practice with others.[73] To support the execution of legal reviews, the ICRC published in 2006 "A Guide to the Legal Review of Weapons, Means and

69. *See* Kenneth Anderson, Daniel Reisner & Matthew C. Waxman, *Adapting the Law of Armed Conflict to Autonomous Weapon Systems*, 90 INTERNATIONAL LAW STUDIES 386, 398 n.27 (2014). For discussion on the customary nature of Article 36, see Natalia Jevglevskaja, *Weapons Review Obligation under Customary International Law*, 94 INTERNATIONAL LAW STUDIES 186 (2018).

70. *See*, *e.g.*, TALLINN MANUAL, *supra* note 65, at 154.

71. WILLIAM H. BOOTHBY, WEAPONS AND THE LAW OF ARMED CONFLICT 344–45 (2d ed. 2016).

72. *Legal Review of New Weapons: Scope of the Obligation and Best Practices*, HUMANITARIAN LAW & POLICY (Oct. 6, 2016), https://blogs.icrc.org/law-and-policy/2016/10/06/legal-review-new-weapons/. For a comprehensive overview of State practice on legal reviews, see COMPENDIUM ON ARTICLE 36 REVIEWS, *supra* note 60. *See also Legal Review of Weapons*, PREMPT, https://www.premt.net/resources/legal-review/ (last visited Feb. 22, 2021).

73. Australia, Sweden, and Norway, for example, have shared their practice in the context of the Group of Government Experts on LAWS. *See*, *e.g.*, Australia, The Australian Article 36 Review Process, U.N. Doc. CCW/GGE.2/2018/WP.6 (Aug. 30, 2018), https://www.unog.ch/80256edd006b8954/(httpassets)/46ca9dabe945fdf9c12582fe00380420/$file/2018_gge+laws_august_working+paper_australia.pdf [hereinafter Australian Article 36 Review Process]. In general, twenty States are known to have in place procedures to conduct the legal review, including Belgium, France, Germany, the Netherlands, the United Kingdom, and the United States. Some States, however, tend not to disclose their practice as it is deemed to be sensitive information with respect to national security. *See* ICRC GUIDE, *supra* note 63, at 5 n.8; *see also* Isabelle Daoust, Robin Coupland & Rikke Ishoey, *New Wars, New Weapons? The Obligation of States to Assess the Legality of Means and Methods of Warfare*, 84 INTERNATIONAL REVIEW OF THE RED CROSS 354 (2002).

Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977" (ICRC Guide).[74]

While the treaty text of Article 36 is not specific, State practice of legal reviews entails certain generalities. Weapons, means and methods of warfare must be assessed in light of the obligations under API, as well as "*any other rule of international law applicable to the High Contracting Party*" according to Article 36 API.[75] The ICRC Guide explains that the reviewer should first assess the existence of specific treaty or customary law provisions outlawing the weapon under review or certain uses thereof.[76] A specific ban or restriction of the use of a weapon, such as the Chemical Weapons Convention's prohibition on the use of chemical weapons,[77] makes any further analysis irrelevant. If no such rules exist, the weapon should be evaluated in light of the fundamental principles of IHL regulating the conduct of hostilities. The vast majority of commentators share the view that there are three relevant principles: (1) the prohibition of using "projectiles and material and methods of warfare of a nature to cause superfluous injuries or unnecessary suffering" of Article 35(2) API[78] and Article 23(e) of the Hague Regulations;[79] (2) the prohibition of employing weapons that—because of their nature—cannot discriminate between military targets and civilians or civilian objects, or be used in a discriminate manner as required by Articles 48 and 51 API;[80] and (3) the prohibition of using weapons that cause widespread, long term and severe damage to the natural environment under Articles 35(3) and 55 API.[81] On the latter principle, disagreement exists as to whether the rule is relevant to every State as a matter of customary law or binds only States party to API as a matter of treaty law.[82]

---

74. ICRC GUIDE, *supra* note 63.

75. API, *supra* note 7, art. 36 (emphasis added).

76. ICRC GUIDE, *supra* note 63, at 11.

77. Convention on the Prohibition of the Development, Production, Stockpiling and Use of Chemical Weapons and on their Destruction art. 1(b), Jan. 13, 1993, 1974 U.N.T.S. 45.

78. API, *supra* note 7, art. 35(2).

79. Hague Regulations, *supra* note 66, art. 23(e).

80. API, *supra* note 7, arts. 48, 51.

81. *Id.* arts. 35(3), 55.

82. According to the *ICRC Guide* on Article 36, for instance, the rule reflects customary law. *See* ICRC GUIDE, *supra* note 63, at 15–16; *Rule 45. Causing Serious Damage to the Natural Environment*, ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule45 (last visited Feb. 22, 2021). For a contrary view, *cf.* BOOTHBY, *supra* note 71, at 351. The ICRC Guide and Boothby agree, instead, on the need to consider "whether there are any

There are further divergent views regarding the applicable normative framework that serves as the reference point for legal reviews. First, it is interesting to note that the ICRC Guide includes the principle of proportionality, i.e., the prohibition of an attack that may be expected to result in excessive civilian harm (deaths, injuries, or damage to civilian objects, or a combination thereof) compared to the concrete and direct military advantage anticipated, among the relevant principles that a legal review should consider. This principle, however, is designed to guide actual conduct on the battlefield. Its application relies on a concrete and specific evaluation made by soldiers or commanders in light of the existing circumstances. Therefore, it is hardly suited as a parameter for legal review. Relevant for the legal review is the weapon's ability to discriminate between lawful and unlawful targets in order to allow its future user to comply with the proportionality principle.[83] Second, it is debated whether legal reviews should consider the weapons' compliance with international human rights law. Only a few States have reported they consider this when assessing weapons that are likely to be employed by armed forces in law enforcement operations.[84] Third, a question remains over the relevance of the "principles of humanity" and the "dictates of public conscience"—the so-called Martens Clause.[85] The ICRC Guide argues that a weapon that is not prohibited or restricted by any specific rule of international law would nonetheless be unlawful if its use would conflict with the principles of humanity and the dictates of public conscience.[86] Most commentators dismiss the Martens Clause as an irrelevant parameter because the principle does not provide normative content on its own.[87] With respect

---

likely future developments in the law of armed conflict that may be expected to affect the weapon" in conducting a legal review. *See* BOOTHBY, *id.* at 348; ICRC GUIDE, *supra* note 63, at 11.

83. *See also* BOOTHBY, *supra* note 71, at 349–50.

84. COMPENDIUM ON ARTICLE 36 REVIEWS, *supra* note 60, at 16.

85. Regulations Respecting the Laws and Customs of War on Land pmbl., annexed to Convention No. II with Respect to the Laws and Customs of War on Land, July 29, 1899, 32 Stat. 1803, T.S. No. 403; API, *supra* note 7, art. 1(2).

86. ICRC GUIDE, *supra* note 63, at 17.

87. *See, e.g.*, BOOTHBY, *supra* note 71, at 351 (quoting YORAM DINSTEIN, THE CONDUCT OF HOSTILITIES UNDER THE LAW OF INTERNATIONAL ARMED CONFLICT 9 (2d ed. 2012); Christopher Greenwood, *Historical Development and Legal Basis*, *in* THE HANDBOOK OF INTERNATIONAL HUMANITARIAN LAW 1, 34–35 (Dieter Fleck ed., 2d ed. 2008)).

to State practice, only Australia is known to take it into consideration, whereas some States simply "keep it in mind" during legal reviews.[88]

When conducting a legal review, States need to consider the characteristics of the given weapon. Reviewers start the analysis by examining how the weapon operates (such as technical specifications, functionality, historical weapons use, extant order of battle, lethal characteristics, and accuracy) as well as weapons data (such as data capture points, data interpretation, and application of data to specifications).[89] Unless it is clear that the weapon cannot comply with IHL in any context or circumstance, they then need to assess how and where the weapon is expected to be used and the "reasonably anticipated effects of employment."[90] In other words, the weapon's characteristics must be analyzed in light of the methods according to which the weapon is intended to be used as well as the intended context.[91] This is necessary as the IHL principles relevant to legal reviews are highly context-dependent, and compliance therewith depends on the environment and circumstances when deployed. Assessing the weapon in light of its expected use allows the reviewer to determine whether the weapon can be lawfully used in a specific setting and/or in combination with certain methods; total or partial restrictions on its use may result.[92]

Yet, there is a nuance to this. Article 36 API requires evaluation of the legality of the weapon in "all or some circumstances,"[93] implying that reviewers must explore all circumstances in which the weapon's use could be unlawful. Regarding the principle of distinction, for instance, a weapon that is

88. COMPENDIUM ON ARTICLE 36 REVIEWS, *supra* note 60, at 3. *See also* Australian Article 36 Review Process, *supra* note 73, at 5 n.20.

89. *See*, *e.g.*, Australian Article 36 Review Process, *supra* note 73, ¶ 6. *See also* U.S. Department of the Air Force, AFI51-401, The Law of War Part 2 (2018) [hereinafter AFI51-401].

90. AFI51-401, *supra* note 89, ¶ 6.1.1.

91. ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES, *supra* note 9, at 22; INTERNATIONAL COMMITTEE OF THE RED CROSS, EXPERT MEETING: AUTONOMOUS WEAPON SYSTEMS, IMPLICATIONS OF INCREASING AUTONOMY IN THE CRITICAL FUNCTIONS OF WEAPONS 23 (2016) [hereinafter AUTONOMOUS WEAPON SYSTEMS, IMPLICATIONS OF INCREASING AUTONOMY]. Furthermore, during the Diplomatic Conference when API was drafted, the rapporteur of Committee III made it clear that "article [36] is intended to require States to analyse whether the employment of a weapon for its normal or expected use would be prohibited under some or all circumstances." COMMENTARY ON THE ADDITIONAL PROTOCOLS, *supra* note 62, ¶ 1469.

92. ICRC GUIDE, *supra* note 63, at 10.

93. API, *supra* note 7, art. 36.

not indiscriminate in itself would be considered unlawful if intended to be used in an indiscriminate manner.[94] There is a limit to this, however. The ICRC *Commentary* on the Additional Protocols notes that States only need to determine "*whether the employment of a weapon for its normal or expected use would be prohibited under some or all circumstances. [It] is not required to foresee or analyze all possible misuses of a weapon, for almost any weapon can be misused in ways that would be prohibited*."[95] Drawing on this, the *Commentary* concludes that the "*obligation only concerns the normal use of the weapon as seen at the time of the evaluation*."[96] Thus, those in charge of the review must have a clear understanding of the technology and sufficient information on the environment and circumstances in which the weapon will be deployed.[97] Therefore, the functioning and effects of the weapons, as well as their use, must be predictable.[98]

V. ASSESSMENT OF COMPLIANCE WITH TARGETING LAW

While fully autonomous weapon systems are not yet a reality, there is a trend towards increasing use of AI for, or in relation to, targeting tasks. Conceptually, autonomy in weapon systems can be categorized according to three different traits, namely (1) the human-machine command-and-control relationship; (2) the sophistication of the machine's decision-making process; and (3) the types of decisions or functions being made autonomous. According to the first trait, systems can be classified based on whether they receive inputs by a human operator to perform their functions in (a) "human-in-the-loop" of the targeting decision, referring to systems that select targets and deliver force upon human command; (b) "human-on-the-loop," capable of selecting targets and delivering force without human interaction but remain under the oversight of humans so that humans retain the power to override the machine's action; and (c) "human-out-of-the-loop," which refers to

94. *See* COMMENTARY ON THE ADDITIONAL PROTOCOLS, *supra* note 62, ¶ 1402.

95. *Id.* ¶ 1469 (emphasis added).

96. *Id.* ¶ 1480 (emphasis added).

97. Jeffrey S. Thurnher, *Examining Autonomous Weapon Systems from a Law of Armed Conflict Perspective*, *in* NEW TECHNOLOGIES AND THE LAW OF ARMED CONFLICT 213, 221 (Hitoshi Nasu & Robert McLaughlin eds., 2014). *See also* Richard Moyes, *Key Elements of Meaningful Human Control*, ARTICLE36 (Apr. 2016), http://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf.

98. Martin Hagström, *Military Applications of Machine Learning and Autonomous Systems*, *in* IMPACT OF ARTIFICIAL INTELLIGENCE, *supra* note 12, at 32, 35.

systems that select targets and deliver force autonomously without humans being able to intervene during the process.[99]

The major difference between manned systems ("human-in-the-loop") and autonomous systems ("human-on-the-loop" and "human-out-of-the-loop") concerns the decision-making process in the targeting cycle. In the most extreme case, it is the system itself that takes targeting decisions based on its observations, perceptions, and evaluations rather than human operators. This is crucial for reviewing the legality of AI systems. For traditional weapons, it is sufficient that the legal review assesses the weapon's technical features, the environment in which it is intended to be deployed, and the intended use. The actual use of the weapon is put in the hands of an operator, who is responsible for ensuring that the IHL rules governing targeting are respected. With AI systems operating autonomously, the role typically performed by the weapon (i.e., releasing force) and that performed by the human operator (i.e., decision-making on the use of force) merge into one unique system. Accordingly, AI-driven weapon systems that conduct targeting decisions need to apply and comply with the entire spectrum of targeting law.[100]

Targeting law refers to those rules under IHL that determine who and what can be targeted in an armed conflict and how. This is also referred to as the rules governing the conduct of hostilities or "Hague Law."[101] Targeting law encompasses three core IHL principles. First, the principle of distinction affirms that parties to the conflict must at all times distinguish between civilians and civilian objects, on the one hand, and legitimate targets (including military objectives), on the other.[102] Second, the principle of

99. *See* PAUL SCHARRE, ROBOTICS ON THE BATTLEFIELD: PART I: RANGE, PERSISTENCE AND DARING 13 (2014), https://s3.us-east-1.amazonaws.com/files.cnas.org/documents/CNAS_RoboticsOnTheBattlefield_Scharre.pdf?mtime=20160906081925&focal=none; VINCENT BOULANIN & MAAIKE VERBRUGGEN, MAPPING THE DEVELOPMENT OF AUTONOMY IN WEAPON SYSTEMS 14 (2017) [hereinafter MAPPING THE DEVELOPMENT OF AUTONOMY]. However, some noted that the notion of "human-out-of-the-loop" may be misleading in so far as no system can possibly be fully "human-free." *See* Patrik Stensson & Anders Jansson, *Autonomous Technology: Source of Confusion: A Model for Explanation and Prediction of Conceptual Shifts*, 57 ERGONOMICS 455 (2014).

100. *See*, *similarly*, Boothby, *Highly Automated and Autonomous Technologies*, *supra* note 59, at 146.

101. *See* STUART CASEY-MASLEN & STEVEN HAINES, HAGUE LAW INTERPRETED: THE CONDUCT OF HOSTILITIES UNDER THE LAW OF ARMED CONFLICT (2018).

102. API, *supra* note 7, arts. 51–52; Protocol Additional to the Geneva Conventions of August 12, 1949, and Relating to the Protection of Victims of Non-International Armed

proportionality prohibits attacks "which 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."[103] Third, the principle of precaution requires that parties to the conflict take "all feasible precautions . . . to avoid, and in any event to minimize, incidental loss of civilian life, injury to civilians and damage to civilian objects."[104]

For legal reviews of weaponized AI, this implies that the system's compliance with the entire range of targeting law must be considered. In addition to the principles already assessed for traditional weapons, as discussed above, the legal review needs to consider the AI system's ability to respect the principle of proportionality, the ability to recognize if a person is *hors de combat*, if he or she has surrendered or is taking direct part in hostilities, and the ability to adopt precautionary measures as required under Article 57 API and customary law, or refer up to those supervising the system and/or planning the attack for them to take the necessary precautionary steps.[105]

This also applies to AI systems that inform humans' decision-making related to targeting or otherwise qualify as means of warfare, although there are limitations. Such systems would need to comply with those rules of targeting law that are relevant to the functions they are entrusted with. For instance, ISR would need to be assessed in light of the principle of distinction. For instance, it would need to be checked to determine if it can properly report that a person is *hors de combat*. While the system would not need to be able to conduct a proportionality assessment itself, it would need to be able to provide the relevant and correct information for assessing the proportionality of an attack and deciding on feasible precautionary measures.

Conflicts art. 13, June 8, 1977, 1125 U.N.T.S. 609. *See also Rule 1. The Principle of Distinction between Civilians and Combatants*, ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule1; *Rule 7. The Principle of Distinction between Civilian Objects and Military Objectives*, ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule7 (both last visited Feb. 22, 2021).

103. API, *supra* note 7, art. 51(5)(b); *Rule 14. Proportionality in Attack*, ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule14 (last visited Feb. 22, 2021).

104. API, *supra* note 7, art. 57(2)(a)–(c); *see also Rule 15. Precautions in Attack*, ICRC, https://ihl-databases.icrc.org/customary-ihl/eng/docs/v1_rul_rule15 (last visited Feb. 22, 2021).

105. Boothby, *Highly Automated and Autonomous Technologies*, *supra* note 59, at 147; BOOTHBY, *supra* note 71, at 348–49.

If the AI system will be used for operations governed by international human rights law, the legal review also needs to consider the system's compliance with the law enforcement paradigm for the use of force. Force can be used only to pursue the legitimate aim of maintaining or restoring public security and law and order.[106] According to the principle of absolute necessity, the use of force must be the last resort. This further implies that only the level of force proportionate to the threat can be used (principle of proportionality) and requires balancing the risk deriving from the individual posing the threat with the potential harm that the use of force may cause to the individual himself and bystanders.[107]

There are procedural consequences arising from the necessity to assess the system's compliance with targeting law and other potentially applicable law. First and foremost, the assessment of an AI system's compliance with targeting law starts during the research and development phase. This is not something peculiar to AI systems. Article 36 API explicitly states that a new weapon's legality shall be duly considered during its study and development.[108] However, for AI systems, this acquires a specific meaning. With traditional weapons, IHL rules operate as external parameters to guide human behavior, whereas, with AI systems, they become part of the system itself.[109] Accordingly, IHL needs to be programmed into the system or "taught" to the system to enable its compliance with IHL's rules.

To this end, developers need to introduce targeting law into the AI system by translating law into standards that an algorithm can understand. They

---

106. Eighth United Nations Congress on the Prevention of Crime and the Treatment of Offenders, *Basic Principles on the Use of Force and Firearms by Law Enforcement Officials*, U.N. Doc. A/CONF.144/28/Rev.1, at 112, ¶¶ 4, 9 (Aug. 27 – Sept. 7, 1990).

107. GLORIA GAGGIOLI, INTERNATIONAL COMMITTEE OF THE RED CROSS, THE USE OF FORCE IN ARMED CONFLICTS INTERPLAY BETWEEN THE CONDUCT OF HOSTILITIES AND LAW ENFORCEMENT PARADIGMS 8 (2012).

108. It provides,

> *In the study*, *development*, acquisition or adoption of a new weapon, means or method of warfare, a High Contracting Party is under an obligation to determine whether its employment would, in some or all circumstances, be prohibited by this Protocol or by any other rule of international law applicable to the High Contracting Party.

API, *supra* note 7, art. 36 (emphasis added.)

109. *See* Joshua A. Kroll et al., *Accountable Algorithms*, 165 UNIVERSITY OF PENNSYLVANIA LAW REVIEW 633, 644 (2017) (quoting EDMUND M. CLARKE JR, ORNA GRUMBERG & DORON PELEG, MODEL CHECKING (1999)).

also need to provide proper data so that the system "learns" relevant IHL rules correctly. This assumes that it is possible to translate the applicable international law into the digital sphere. Indeed, lawyers recognize that there is "an increasing need for law in algorithmic forms."[110] Its technical feasibility is not yet guaranteed, however. The main challenges are the translation of the nature of the law into the algorithm's language and the associated risks of unintended consequences.

As AI systems operate (and learn) based on the data they receive, data becomes central in developing the system and determining how it will operate. Experts will need to ensure that the system is trained with appropriate data or, if the system collects data and learns autonomously, programmed such that it will collect only appropriate data and properly use it. This means that the system needs to be trained with a focus on the environment in which it is to be deployed and in circumstances that are representative of that environment.[111] If a machine learning system entrusted with recognizing legitimate targets is trained only on people of a certain nationality or ethnicity, the system may associate that nationality or ethnicity with being an enemy and thereby possibly confuse civilians or persons taking no direct part in hostilities with legitimate targets, for instance.

With data becoming the key determinant of an AI system's output, it follows that selection and revision of such data need to be at the heart of the legal review of an AI system. As a result, legal advisers with experience in conducting legal reviews must engage during the design and development phase to support computer scientists and engineers in developing and programming systems that conform to relevant legal principles.[112] Their legal

110. Lisa Shay et al., *Do Robots Dream of Electric Law? An Experiment in the Law as Algorithm*, *in* ROBOT LAW 274, 298 (Ryan Calo, A. Michael Froomkin & Ian Kerr eds., 2016). For a further discussion, see notably Deeks, *supra* note 28; Laurie Blank, *New Technologies and the Interplay between Certainty and Reasonableness*, *in* COMPLEX BATTLESPACES: THE LAW OF ARMED CONFLICT AND THE DYNAMICS OF MODERN WARFARE 317 (Winston S. Williams & Christopher M. Ford eds., 2019).

111. MICROSOFT CORPORATION, THE FUTURE COMPUTED: ARTIFICIAL INTELLIGENCE AND ITS ROLE IN SOCIETY 64 (2018). This requirement is not specific to military AI or machine learning applications. Yet, it is worth mentioning that, to date, efforts to put it in practice have regularly failed. E-mail from Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe, to the authors (Sept. 28, 2020, 23:35 CST) (on file with the authors).

112. On the integration of technical knowledge with legal expertise of commanders and judge advocates, see Annemarie Vazquez, *Laws and Lawyers: The LOAC Needs Judge Advocates*

knowledge would contribute to transforming relevant legal parameters into understandable standards and selecting representative data to allow the system to learn IHL correctly.[113] Once data is fed into the system, legal advisers could also evaluate the results from a legal standpoint by taking part in the testing process. The technical expertise, combined with the legal expertise, would ultimately help ensure that the system performs safely, free from biases, and lawfully.[114] This process can be equated to an "anticipated" legal review, which, rather than being carried out on the system as a final product, has mostly data as its object.

Accordingly, when comparing the legal review of traditional weapons with the legal review of AI-driven or -supported systems, the first differences are that targeting law must be considered, the legal review must be conducted at an earlier stage in the development process of the system, and the appropriateness of data needs to be considered. This, however, still builds on the assumption that the system's behavior at the time of the legal review will be the same after the legal review, i.e., the behavior does not evolve in ways that are inconsistent with original expectations.

VI. THE PREDICTABILITY PROBLEM

When humans operate traditional weapons, the weapon's anticipated use is based on expectations of the operator's behavior in combat. This depends to a large extent on the operator's capacities, training, and experience. This is also guided by standard operating procedures (SOPs), rules of engagement (ROE), orders, and other regulations and administrative measures concerning the operator's decision-making. During a legal review of a weapon, these elements must be taken into account in a general and abstract manner. Future human behavior can be estimated based on prior experience with humans in battle. While this does not allow conclusions with full certainty, it does allow certain levels of confidence. In addition, the legal reviewer knows that the person operating the weapon will have full responsibility and accountability over the weapon's use, as every combatant is bound to apply IHL. This ensures a high level of predictability of the future use of the

at the Design Table for Lethal Autonomous Weapons, and We Can Start Now, 228 MILITARY LAW REVIEW 89 (2020).

113. Questions would encompass, for example, whether the data provided are enough to allow the algorithm to recognize that a person is *hors de combat*, a combatant, or a civilian taking direct part in hostilities.

114. MICROSOFT CORPORATION, *supra* note 111, at 63.

weapon. For AI systems, this is different. The nature of AI applications introduces inherent uncertainty as to how systems will behave once deployed and respond to changing and complex environments. Their unpredictability is of a different nature and degree depending on the AI techniques employed, namely hand-coded programming or machine learning. This technological distinction is relevant for adapting legal reviews to AI systems.

A. Hand-coded Programming

The first and most traditional approach to AI is hand-coded programming.[115] This approach relies on the elaboration by a programmer of a model of the world, including the rules of logic governing the relationships therein, which is then crafted into the system to allow it to operate autonomously.[116] Programmers are required to research how the world works and create a model of the universe that describes the environment where the system is intended to be employed. Because the model, or source code, is crafted by human programmers, the system's inner functioning is readable and understandable by humans. Accordingly, by providing a system with certain inputs, it is possible to observe the system's response and understand what process the system has followed to reach a specific outcome.[117]

A major limitation is that the handcrafted approach is only suited if the operational environment can be reduced to clear mathematical rules. Armed conflicts are typically subject to constant changes and are not predictable. The process of transforming this operational environment into the code can therefore be extremely challenging, if not impossible.[118] This has the consequence that handcrafted AI is only operationalizable in limited scenarios of armed conflicts. Currently, such systems have been deployed to undertake

115. It is noteworthy that technological developments have evolved so far that some experts consider hand-coded programming as no longer reflecting AI proper. According to them, the notion of AI seems to have shifted toward that of machine learning/reinforcement learning and algorithms operating with big amounts of data.

116. Boulanin, *Artificial Intelligence*, *supra* note 12, at 19.

117. *AI Next Campaign*, *supra* note 23.

118. Notably, when it comes to complex realities, it might be impossible to provide an algorithm with the rules of logic it needs to work as desired. *See* Hagström, *supra* note 98, at 36; Leon Kester, *Mapping Autonomy, in* LETHAL AUTONOMOUS WEAPONS SYSTEMS: TECHNOLOGY, DEFINITION, ETHICS, LAW & SECURITY 196, 199–200 (Robin Geiss ed., 2017), https://www.auswaertiges-amt.de/blob/610608/5f26c2e0826db0d000072441fdeaa8ba/abruestung-laws-data.pdf.

military tasks underwater, which represent uncluttered and highly predictable scenarios.[119]

For legal reviews, the consequence is that a reviewer can assess the system's legality with regard to the specific settings in which it has been tested. Because handcrafted systems operate under a rigid framework of parameters crafted into it, a reviewer can rely on the fact that the system's performance as observed during the review would remain unchanged once deployed and subjected to similar inputs in the real world. In addition, handcrafted systems' suitability for relatively simple environments makes it easier to reproduce such environments and makes testing less cumbersome because of the limited set of inputs characterizing such environments. By testing the system in the context of its end-task, the reviewer can gain direct and precise evidence of the system's lawful or unlawful functioning in those specific settings.[120]

Yet, it remains impossible to test a system for every possible input it could face at the operational stage. Therefore only a limited number of outcomes could ever be observed and evaluated.[121] Generally speaking, for inputs that have not been tested, the AI system remains unpredictable and likely to fail if confronted with situations that differ from those that have been verified.[122] This is often called "brittleness."[123] Since a legal review

119. For example, when it comes to navigating functions, it is easier to create systems that can perform such tasks in the air or underwater because compared to land they present a limited number of possible obstacles and they are governed by laws of physics that make them representable in mathematical terms. The Sea Hunter, a maritime autonomous system to hunt for nuclear-powered ballistic missile submarines, represents one such example. *See* MAPPING THE DEVELOPMENT OF AUTONOMY, *supra* note 99, at 14, 23. *See also* Michael W. Meier, *Emerging Technologies and the Principle of Distinction: A Further Blurring of the Lines between Combatants and Civilians?*, *in* IMPACT OF EMERGING TECHNOLOGIES, *supra* note 49, at 211, 225; PETER LAYTON, ALGORITHMIC WARFARE: APPLYING ARTIFICIAL INTELLIGENCE TO WARFIGHTING 64 (2018).

120. Finale Doshi-Velez & Been Kim, *Towards A Rigorous Science of Interpretable Machine Learning* 4–5 (2017), https://arxiv.org/pdf/1702.08608.pdf .

121. Kroll et al., *supra* note 109, at 633.

122. Nevertheless, it might be possible that a well-trained system may execute its tasks effectively without ever seeing the input before, depending on the offline learning or feedback during online learning.

123. For example, autopilot systems that help airplane pilots navigate and reduce human error follow their programming precisely every time, but they are brittle when used outside of their intended operating environment. Scharre & Horowitz, *supra* note 30. *See also* Sauer, *supra* note 14, at 85.

cannot test and ensure the legality of the weapons' functioning in such circumstances, the consequence of the legal review would be to impose clear limitations on the weapon's deployment. Concretely, this would mean to specify that the weapons' use would only be possible within the tested context or to prescribe a certain behavior in conformity with IHL, such as withholding kinetic effects against a person or object when facing untested inputs. Indeed, any action outside such a framework could not have been foreseen and thus potentially illegal.

B. Machine Learning

The second type of AI is machine learning. Machine learning refers to a process that allows a system to learn by "discover[ing] correlations between variables in a dataset, often to make predictions or estimates of some outcome."[124] Through machine learning, a system is "trained" using large amounts of data,[125] from which it identifies correlations, builds a representation of the world, and ultimately learns how to perform a task.[126] Given the system's ability to learn for itself, it does not require explicit programming.[127] Instead, those who develop the system need to create a structure that allows the system to learn and adapt to changing situations.[128] They also need to provide the algorithm with a great amount of properly selected data concerning the operating environment.

Learning capabilities imply that, contrary to handcrafted systems, the system changes "its structure, program, or data (based on its inputs or in response to external information) in such a manner that its expected future

124. David Lehr & Paul Ohm, *Playing with the Data: What Legal Scholars Should Learn About Machine Learning*, 51 UC DAVIS LAW REVIEW 653, 671 (2017).

125. This is most notably true for machine learning relying on deep learning artificial neural networks.

126. Michael Copeland, *What's the Difference Between Artificial Intelligence, Machine Learning and Deep Learning?*, NVIDIA (July 29, 2016), https://blogs.nvidia.com/blog/2016/07/29/whats-difference-artificial-intelligence-machine-learning-deep-learning-ai/.

127. This is what makes machine learning suitable for complex cognitive tasks such as target recognition or, beyond the military sphere, translation of a language to another. *See*, *e.g.*, Lewis-Kraus, *supra* note 15; DEFENSE INNOVATION BOARD, *supra* note 27, at 46; Hagström, *supra* note 98, at 36.

128. 1 OFFICE OF CHIEF SCIENTIST, U.S. AIR FORCE, AUTONOMOUS HORIZONS: SYSTEM AUTONOMY IN THE AIR FORCE – A PATH TO THE FUTURE: HUMAN-AUTONOMY TEAMING § 5.2 (June 2015), https://www.af.mil/Portals/1/documents/SECAF/AutonomousHorizons.pdf.

performance improves."[129] As an example, if employed in conflict settings, a machine learning system may develop its own criteria to apply a set of rules through observations made on the battlefield.[130] The learning process may occur in different manners: supervised learning, unsupervised learning, and reinforcement learning. In supervised learning, the system is trained through a combination of input instances and output labels and then trained to generalize a function. Through unsupervised learning, the system is provided only with inputs, and it finds structures and features in the data. In reinforcement learning, the learning process occurs through trial and error. The future action is chosen either because it maximizes a future reward[131] or because the system is seeking to learn something new.[132] Such learning processes can occur either offline or online. In the former case, the algorithm is trained while being developed. In the latter, the system learns—or keeps learning—during the deployment phase and adapts to the environment.[133] Due to these technical features, machine learning systems are well-suited to operate in complex scenarios where hand-coded programming would fail. In the military domain, machine learning is already employed to perform target recognition tasks, for instance.[134]

Machine learning systems' increased flexibility and autonomy at the operational stage lead to major challenges for their legal assessment, however. The first problem derives from machine learning's reliance on training data to learn. This has been paraphrased as "the intelligence is in the data, not the

129. NILS J. NILSSON, INTRODUCTION TO MACHINE LEARNING: AN EARLY DRAFT OF A PROPOSED TEXTBOOK 1 (1998). The improvement of a machine learning system is measured in some metric of performance that developers choose. However, improvements based on such a metric do not necessarily reflect good performance in undertaking the task as it would be defined to a human operator. E-mail from Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe, to the authors (Sept. 28, 2020, 23:35 CST) (on file with the authors).

130. BOOTHBY, *supra* note 71, at 251.

131. The system does not necessarily learn to maximize immediate rewards, but it can learn to perform actions that will lead to rewards later.

132. PERRY VAN WESEL & ALWYN E. GOODLOE, CHALLENGES IN THE VERIFICATION OF REINFORCEMENT LEARNING ALGORITHMS 5–6 (2017).

133. *Id.*

134. *See*, *e.g.*, Jon Barker, *From the Frontline: How Deep Learning Plays Critical Role in Military Problem-Solving*, NVIDIA (June 29, 2016), https://blogs.nvidia.com/blog/2016/06/29/deep-learning-6/.

algorithm."[135] As mentioned above, the systems learn and operate based on data. Therefore, it is essential that data be representative of the reality where the system is intended to operate to avert unexpected bias or serious flaws in the system's functioning.[136] In addition, because of the complex scenarios where these systems are typically employed, the amount of inputs they could face is exponentially higher than that faced by handcrafted systems. Since these inputs can be potential sources for new learning and change the system's operating parameters, this would cause greater uncertainty as to what the system would learn and how it would react to inputs.

The other challenge is directly linked to how machine learning systems reach their conclusions given certain inputs, also known as the "black box" or opacity problem. Machine learning algorithms, notably those relying on deep learning artificial neural networks, operate like a "black box" in the sense that while inputs and outputs of their functioning are observable, the inner process by which it reaches a specific outcome cannot be deconstructed and understood.[137] This means that the system lacks the ability to give reasons for its estimations or decisions—it lacks explainability.[138] When a system's learning process is limited to the offline phase, it will no longer change once a military makes the system operational. Although this might represent an advantage in terms of the foreseeability of a future system's behaviors, it does not necessarily imply that the system acts deterministically, i.e., produces the same outputs if subjected to the same inputs.[139] Yet even when deterministic AI systems are concerned, predictability remains limited. It would be unrealistic to expect that such systems receive exactly the same inputs they received during the pre-deployment assessment in real-life

135. BRIAN A. HAUGH, DAVID A. SPARROW & DAVID M. TATE, THE STATUS OF TEST, EVALUATION, VERIFICATION, AND VALIDATION (TEV&V) OF AUTONOMOUS SYSTEMS ¶ 10, at 2-3 (2018).

136. *See generally* William Kruskal & Federick Mosteller, *Representative Sampling, III: The Current Statistical Literature*, 47 INTERNATIONAL STATISTICAL REVIEW 245 (1979).

137. Anjanette H. Raymond, Emma Young & Scott J. Shackelford, *Building a Better HAL 9000: Algorithms, the Market, and the Need to Prevent the Engraining of Bias*, 15 NORTHWESTERN JOURNAL OF TECHNOLOGY & INTELLECTUAL PROPERTY 215, 221 (2018).

138. Lehr & Ohm, *supra* note 124, at 706 (quoting Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 BIG DATA & SOCIETY 1, 1–2, https://journals.sagepub.com/doi/pdf/10.1177/2053951715622512 (2016)).

139. VAN WESEL & GOODLOE, *supra* note 132, at 13.

situations.[140] As illustrated by examples with adversarial attacks, very small changes in the inputs can lead to drastic changes in the outputs despite the model's deterministic nature.[141] This challenge is exacerbated for online machine learning. Although such systems can be subjected to specific inputs during a test, it remains impossible to generalize from the observed input-output correlations and foresee how the system would react when confronted with similar inputs in the real world.[142] In other words, depending on what they learn during the operational phase and environmental conditions, non-deterministic algorithms can produce different outputs even when subjected to identical inputs.[143]

The complexity of such models highlights the tradeoff between autonomy and predictability. In practical terms, full predictability will not be an achievable state. It follows that such systems would never pass the legal review if full predictability is set as a precondition for an AI system's compliance with IHL.[144] Setting such a bar would also require more than what is expected when humans are in the targeting decision-making loop. Members of the armed forces are entrusted with targeting tasks and assessments even if they may behave unexpectedly and/or contrary to what has been ordered or is legally required. Yet, commanders still deploy them. For AI systems, rather than setting an unachievable standard, it is more realistic to accept that a certain degree of unpredictability is an intrinsic trait thereof. Indeed, the

---

140. Telephone Interview with Benjamin Schumeg, Tarek Abulmagd, Adam Hilburn, Adam Hoxha, Newman Hsiao, Ryan Olsen, Katy Perez, Gagan Singh, and Carl Valianti, United States Army Futures Command, Software Qualification Branch (Nov. 5, 2020).

141. *See*, *e.g.*, Carl Velasco, *Artificial Intelligence Thinks This Turtle Is a Gun, Highlighting A Major Problem with Object Recognition Technology*, TECH TIMES (Nov. 3, 2017), http://www.techtimes.com/articles/215163/20171103/artificial-intelligence-thinks-this-turtle-is-a-gun-highlighting-a-major-problem-with- object- recognition- technology.htm; *see also* Knight*, supra* note 17.

142. For an analysis of the predictability problem from a *jus ad bellum* perspective, see Ashley Deeks, Noam Lubell & Daragh Murray, *Machine Learning, Artificial Intelligence, and the Use of Force by States*, 10 JOURNAL OF NATIONAL SECURITY LAW & POLICY 1, 21 (2019).

143. Tim Menzies & Charles Pecheur, *Verification and Validation and Artificial Intelligence, Foundations 02: A V&V Workshop* 39 (Sept. 15, 2002); *See also* MICHÈLE A. FLOURNOY, AVRIL HAINES & GABRIELLE CHEFITZ, BUILDING TRUST THROUGH TESTING: ADAPTING DOD'S TEST & EVALUATION, VALIDATION & VERIFICATION (TEVV) ENTERPRISE FOR MACHINE LEARNING SYSTEMS, INCLUDING DEEP LEARNING SYSTEMS 8 (2020), https://cset.georgetown.edu/wp-content/uploads/Building-Trust-Through-Testing.pdf.

144. *See*, *e.g.*, AUTONOMOUS WEAPON SYSTEMS, IMPLICATIONS OF INCREASING AUTONOMY , *supra* note 91, at 9.

use of AI may transform IHL's standards from the "reasonableness" associated with human decision-making towards probabilistic levels of "certainty."[145]

For legal reviews, this means that different levels of predictability could be determined depending on the function attributed to the system. Notably, if it is to undertake critical functions related to targeting—and on the environment in which it is to be deployed—the more critical the tasks, the higher the level of predictability that needs to be satisfied. For instance, a system performing targeting functions would need to satisfy a very high level of confidence regarding the lawfulness of a target before being able to fire, even if teamed with a human operator.[146] The level of acceptable predictability set for each type of system-scenario would then become the parameter against which to assess the algorithm as a precondition to passing the legal review. The alternative would be to deploy the system in very limited operational circumstances, thus limiting its potential, or setting some safety mitigations, which likewise might limit its functions to non-critical ones.[147] Given such tradeoffs, it is crucial not to undermine existing standards under IHL. Yet, from a practical perspective, evaluating if the given standards are met by an AI system is first and foremost a technical issue.

VII. CONGRUENCE OF VERIFICATION AND VALIDATION WITH THE LEGAL REVIEW

After a new weapon or weapon system is developed, it needs to undergo technical testing and approval before it can be legally reviewed and eventually deployed. This is accomplished through the process of "verification and validation" (V&V). One of the best-established V&V processes for weapons is that of the U.S. DoD. V&V is a technical process involving different expertise, including engineering, mathematics, and computer science, that leads to technical certification of the system.[148] "Verification" consists of

145. Blank, *supra* note 110.

146. It is noteworthy that an autonomous system would likely be composed of multiple subsystems. As such, the algorithm that makes decisions may be compliant but not the one that makes the recognition. It is important that such potential for error does not go unnoticed.

147. *See also* FLOURNOY, HAINES & CHEFITZ, *supra* note 143, at 22.

148. In the United States, for instance, the Army's process is conducted by a specialized unit, the Army Evaluation Branch. Private contractors are free to conduct and provide their own validation and verification outcomes, yet the Army Evaluation Branch never relies

mathematically "determining that [a system] accurately represents the developer's conceptual description and specifications"[149] and allows developers to evaluate "the extent to which [the system has] been developed using sound and established software-engineering techniques."[150] It is applied at each stage of the life cycle management process to ensure that the inputs and outputs are implemented accurately and properly.[151] Verification basically answers the question: does the system do what the programmers said? In other words, it investigates whether the system has been built in the right manner.[152] "Validation" is the process of determining the extent to which the system is an accurate representation of the real world from the perspective of its intended use and assuring that the system meets the needs of those who will utilize it.[153] Informally speaking, it means asking whether the right model has been built.[154]

V&V is typically combined with an additional technical process consisting of weapons "testing and evaluation" (T&E), usually occurring at an earlier stage. T&E is the primary means to ensure that the system will actually

completely on such results. Telephone Interview with Benjamin Schumeg, Tarek Abulmagd, Adam Hilburn, Adam Hoxha, Newman Hsiao, Ryan Olsen, Katy Perez, Gagan Singh, and Carl Valianti, United States Army Futures Command, Software Qualification Branch (Nov. 5, 2020).

149. U.S. DEPARTMENT OF DEFENSE, TEST AND EVALUATION MANAGEMENT GUIDE 220 (6th ed. 2012), https://www.dau.edu/tools/Lists/DAUTools/Attachments/148/Test%20and%20Evaluation%20Management%20Guide,%20December%202012,%206th%20Edition%20-v1.pdf [hereinafter TEST AND EVALUATION MANAGEMENT GUIDE].

150. HEADQUARTERS, U.S. DEPARTMENT OF THE ARMY, PAMPHLET 5-11, VERIFICATION, VALIDATION, AND ACCREDITATION OF ARMY MODELS AND SIMULATIONS ¶ 3-2 at 24 (1999), https://armypubs.army.mil/epubs/DR_pubs/DR_a/pdf/web/p5_11.pdf [hereinafter VERIFICATION, VALIDATION, AND ACCREDITATION OF ARMY MODELS AND SIMULATIONS].

151. Fei Liu, Ming Yang & Peng Shi, *Verification and Validation of Fuzzy Rules-Based Human Behavior Models*, *in* 7TH INTERNATIONAL CONFERENCE ON SYSTEMS SIMULATION AND SCIENTIFIC COMPUTING 813 (2008).

152. Dean S. Hartley III, *Verification & Validation in Military Simulations*, *in* PROCEEDINGS OF THE 1997 WINTER SIMULATION CONFERENCE 925 (Sigrun Andradóttir et al. eds., 1997).

153. BEN H. THACKER ET AL., CONCEPTS OF MODEL VERIFICATION AND VALIDATION 2 (Charmian Schalle ed., 2004); Hartley III, *supra* note 152; DEFENSE INNOVATION BOARD, *supra* note 27, at 47.

154. Hartley III, *supra* note 152.

perform its intended functions in its intended environment.[155] It aims to demonstrate "the feasibility of conceptual approaches, evaluate design risk, identify design alternatives, compare and analyze tradeoffs, and estimate satisfaction of operational requirements."[156] The phase covers both the development stage of the system as well as the operational one, in which the developers investigate the system's operational effectiveness, suitability, and survivability. According to the U.S. DoD,

> test and evaluation shall be structured to provide essential information to decision makers, assess attainment of technical performance parameters, and determine whether systems are operationally effective, suitable, survivable, and safe for intended use. The conduct of test and evaluation, integrated with modeling and simulation, shall facilitate learning, assess technology maturity and interoperability, facilitate integration into fielded forces, and confirm performance against documented capability needs and adversary capabilities as described in the system threat assessment.[157]

When it comes to AI systems operating through machine learning, the T&E is directly linked to the system training process. Most of the algorithms learn thanks to two sets of data. On the one hand, the training data set is fed to the algorithm. On the other hand, the testing data set allows the developer to test and evaluate what the algorithm has learned.[158]

States' efforts towards integrating autonomy within their military capabilities lead them to focus on the above-mentioned processes. In fact, one of the major obstacles to increasing military autonomy is the lack of suitable V&V methods for AI. Given the deterministic nature of model-based AI systems, current techniques for verifying and validating can be easily applied thereto.[159] The same holds true for machine learning with offline capabilities.

---

155. BERNARD FOX ET AL., RAND CORPORATION TEST AND EVALUATION TRENDS AND COSTS FOR AIRCRAFT AND GUIDED WEAPONS at xv (2004).

156. TEST AND EVALUATION MANAGEMENT GUIDE, *supra* note 149, ¶ 2.1 at 23.

157. *Id.* ¶ 2.3 at 35.

158. VAN WESEL & GOODLOE, *supra* note 132, at 6. Ideally, the test set is only used once to evaluate the performance and no further tuning should be done in the system. Yet, many developers do use the test set several times (a practice known as "double dipping"). This leads to optimistic evaluations of performance and poor generalization capabilities. E-mail from Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe, to the authors (Sept. 28, 2020, 23:35 CST) (on file with the authors).

159. Liu, Yang & Shi, *supra* note 151, at 813.

Once trained, the system's learning capabilities are frozen, and the algorithm does not receive any more inputs nor change its structure. Although it might not be easy to analyze the results and identify appropriate use cases to test the system, conventional V&V approaches remain valid and applicable.[160] For machine learning with online learning capabilities, verification becomes more complex due to the non-deterministic and adaptive nature of such algorithms. It might be impossible to verify every version of a machine learning system (weapon or not) given its potential continuous change.[161]

This challenge has not gone unnoticed by States that intend to leverage AI's potential. The United States highlights that the lack of appropriate V&V methods and effective T&E of autonomous systems "prevents all but relatively low levels of autonomy from being certified for use."[162] This underlines the need for States to develop mechanisms of V&V of AI that ensure the systems can be trusted.[163] Confronted with this challenge, experts are working on identifying methods to conduct effective verification of machine learning systems. One suggestion that is gaining growing consensus is to conduct "runtime verifications" on online machine learning systems in order to keep up with the system adaptations to the environment.[164] This would shift the focus toward a continuous evaluation during the full life cycle of the system, first and foremost the operational phase. The system would be certified prior to deployment to attest its suitability for use in a limited set of scenarios, while the incompleteness of the processes and potential unintended scenarios would be acknowledged. Developers should then continuously monitor performance and report to regulators to allow re-evaluation and take corrective measures if necessary.

By doing so, the system is assessed in real operating conditions, and violations of its properties and specifications are detected and addressed while the system is running. It would also solve two problems at the same time. First, continuous verification during system operation would ensure that although its unpredictability is accepted, this would not lead to violations of the system's specifications and applicable law. Second, it would provide

160. VAN WESEL & GOODLOE, *supra* note 132, at 12.

161. Menzies, *supra* note 143, at 41, 60.

162. DAHM, *supra* note 36. S*ee also* Fil Macias, *The Test and Evaluation of Unmanned and Autonomous Systems*, 29 ITEA JOURNAL 388 (2008).

163. ZACHARIAS, *supra* note 25, at 253, 277.

164. DEFENSE INNOVATION BOARD, *supra* note 27, at 36; VAN WESEL & GOODLOE, *supra* note 132, at 12.

those responsible for system functioning with a tool to retain control over it. This is relevant from an operational standpoint since no commander would entrust a system operating outside its sphere of control with critical functions. Yet there is a nuance: although runtime verification can be suited to machine learning with online capabilities, to date it can be effectively applied only when it is possible to specify criteria to constrain a machine learning adaptive system.[165]

Further options are variants of a verification method known as "model checking."[166] Model checking is an algorithmic (fully automated) method for determining if a model of a system satisfies a correctness specification or property.[167] In order to determine whether the specification is satisfied, model checking allows the developer to undertake an exhaustive exploration of a system's achievable states, that is, all the possible executions of the system. If a state that violates a correctness property is found, a counter-example is produced to demonstrate the error.[168] The problem with this verification technique is that it is applicable only to finite-state systems,[169] whereas the number of states in machine learning systems, particularly those working through deep neural networks, is enormous, if not infinite.[170] In such a case, traditional model checking would be unsuited because it is impossible to explore all possible states and transitions of the system. Therefore, experts have proposed techniques to make model checking suitable to nonfinite state

165. VAN WESEL & GOODLOE, *supra* note 132, at 12.

166. DEFENSE INNOVATION BOARD, *supra* note 27, at 35.

167. VAN WESEL & GOODLOE, *supra* note 132, at 12; Vijay D'Silvay, Daniel Kroening & Georg Weissenbacher, *A Survey of Automated Techniques for Formal Software Verification*, 27 IEEE TCAD 1165 10 (2008). Although this seems a promising approach, there might be limitations on the scalability both in the case of nonfinite states as well as when applied to complex problems. Indeed, the challenges of defining the correctness specification may be similar to those that confront scientists and engineers in designing the AI system itself. E-mail from Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe, to the authors (Sept. 28, 2020, 23:35 CST) (on file with the authors).

168. D'Silvay, Kroening & Weissenbacher, *supra* note 167, at 5; Charles Pecheur & Reid Simmons, *From Livingstone to SMV: Formal Verification for Autonomous Systems*, NTRS 3 (Jan. 2, 2000), https://ntrs.nasa.gov/citations/20010081205.

169. Edmund M. Clarke, William Klieber, Miloš Nováćek & Paolo Zuliani, *Model Checking and the State Explosion Problem*, *in* TOOLS FOR PRACTICAL SOFTWARE VERIFICATION 1, 10 (Bertrand Meyer & Martin Nordio eds., 2012).

170. Martin Kot, *The State Explosion Problem* (2003), http://www.cs.vsb.cz/kot/down/Texts/StateSpace.pdf.

models that would allow the verification of machine learning with adaptive functions.[171] Yet, it is worth mentioning that to date these new approaches to model checking and runtime verification remain fields of research in the early phases of real-life validation.[172]

Certainly, effective verification of machine learning would benefit from increasing the system's explainability. Explainable AI has been specifically identified as a way to overcome the problem of the "black box" and related unpredictability of machine learning. Explainable AI enables AI systems to give explanations for their outcome or prediction that are understandable by humans. It enables users to "understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" without giving up the system's accuracy.[173] This facilitates verification since runtime verification could focus on these explanations as indicators of certain properties of the system.[174] So far, however, explainable AI also remains a subject of research, and its level of adoption limited.

This leads to another important aspect to consider, namely the need to automate the process of verifying the system's robustness. Originally, software certification was mostly manual and conducted by humans. Nowadays, verification processes can benefit from the support of automation. Manual certification mechanisms, besides being time consuming and costly, can result in fouled evaluations because reviewers introduce their own sets of expertise, experiences, and biases. Hence, DARPA developed an automated rapid software certification program to allow an automatic assessment of software evidence. This permits certifiers to rapidly determine that the

171. In this regard, major advancements are symbolic model checking, partial order reduction, counter-example-guided abstraction refinement, and bounded model checking. For further insights on such techniques, see Clarke, Klieber, Nováćek & Zuliani, *supra* note 169, at 8-29; D'Silvay, Kroening & Weissenbacher, *supra* note 167, at 6–11.

172. Experts working on AI are proceeding by trial and error and draw from other fields where potentially suitable techniques have been used. For instance, redundancy is a technique already in use in the aero-spatial domain. Telephone Interview with Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe (Nov. 11, 2020).

173. Matt Turek, *Explainable Artificial Intelligence (XAI)*, DARPA, https://www.darpa.mil/program/explainable-artificial-intelligence (last visited on Feb. 22, 2021).

174. VAN WESEL & GOODLOE, *supra* note 132, at 20.

system risk is acceptable and move away from document-based engineering processes.[175]

This state-of-the-art assessment, along with foreseeable developments in AI certification, indicates that the solutions to the predictability problem are found in the technical field, not the legal. This, however, is indicative of a fundamental consequence for the legal review of AI-driven systems: the technical and legal assessment conflate into one single assessment—the legal review becomes congruent with V&V. Traditionally, V&V procedures are separate from, though functional with, legal reviews.[176] The U.S. DoD directive "The Defense Acquisition System," for instance, examines the technical and legal assessments as two distinct steps of the weapon and system acquisition process.[177] Weapon testing and technical assessment provide empirical evidence of the weapon performance on which militaries and legal experts can base their legal review.[178] Data supporting the review include the results of any tests on weapon accuracy, reliability, performance, wounds, failure rates, or other relevant matters.[179]

Yet, for AI systems, technical verification can work as both a technical *and* legal assessment. The reason is that, as discussed above, targeting law is transformed into technical parameters and embedded into the system during the designing phase. Once the legal standards are learned by the system, compliance with those standards is a technical task. Regarding the verification process—namely ensuring that the system matches the developer's conceptual description and specifications—verifying an AI weapon system allows the State to check whether the system matches technical specifications regarding targeting law. Likewise, concerning validation, the assessment of

175. *Expediting Software Certification for Military Systems, Platforms*, DARPA (Mar. 5, 2019), https://www.darpa.mil/news-events/2019-05-03.

176. ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES, *supra* note 9, at 24. According to the U.S. Air Force instruction, for the purpose of conducting the legal review, Air Force personnel shall provide, among others, "the reasonably anticipated effects of employment, to include all tests, computer modeling, laboratory studies, and other technical analysis and results that contribute to the assessment of reasonably anticipated effects." AFI51-401, *supra* note 89, at 10.

177. *See* U.S. Department of Defense, DoD Directive 5000.01, The Defense Acquisition System 9 (2020), https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodd/500001p.pdf [hereinafter DoD Directive 5000.01 (2020)].

178. ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES, *supra* note 9, at 23. *See also* TALLINN MANUAL, *supra* note 65, at 129, ¶ 11.

179. BOOTHBY, *supra* note 71, at 352.

whether an AI system meets the needs of those utilizing it allows the State to assess the system regarding its expected use, duly considering the environment in which it is to be deployed. These two steps indeed satisfy all that IHL requires from legal reviews of weapons and means of warfare. Accordingly, the V&V and legal review become one.

A practical consequence of this is that legal knowledge and experience will need to be integrated with technical expertise in the V&V process. First, this should entail the participation of legal experts in the V&V procedure concerning data selection, as observed above. Technical experts could oversee the procedures, whereas legal experts would assess whether the behaviors and outcomes observed during such phases are compliant with targeting law. To this end, the peer review mechanism established in the United States as part of the V&V process[180] could serve as a basis for such integration of legal expertise. This is also in line with the collegiality principle already existing in some legal review committees, which intends to include every necessary expertise for the assessment of the lawfulness of a weapon.[181] Second, cross-fertilization between legal and technical spheres implies that V&V experts need to be familiar with the legal principles and rules applicable to targeting, at least those regulating the system's tasks. Likewise, legal experts would need a basic understanding of the technology and technological issues involved.[182] This would compensate for lawyers' potential lack of understanding of the underlying technology, which may arise with complex AI systems.[183] Ultimately, this reciprocal exchange of knowledge would allow bridge-building and favor understanding and cooperation between the technical and legal sphere. Such an integrated process would reflect the holistic (technical and legal) assessment that such systems require.

---

180. *See* VERIFICATION, VALIDATION, AND ACCREDITATION OF ARMY MODELS AND SIMULATIONS, *supra* note 150, at 30.

181. BOOTHBY, *supra* note 71, at 354. Moreover, collaboration and integration of knowledge already characterize other processes, such as acquisition processes by the U.S. DoD. Accordingly, "[t]eaming among warfighters, users, developers, acquirers, technologists, testers, budgeters, and sustainers shall begin during capability needs definition." *See* U.S. Department of Defense, DoD Directive 5000.01, The Defense Acquisition System ¶ E1.1.2 (2003, incorporating Change 2, Aug. 31, 2018), https://www.acq.osd.mil/jrac/docs/DoD-Directive-5000.01.pdf (superseded by DoD Directive 5000.01 (2020), *supra* note 177).

182. *See also* FLOURNOY, HAINES & CHEFITZ, *supra* note 143, at 26; DAVID LESLIE, UNDERSTANDING ARTIFICIAL INTELLIGENCE ETHICS AND SAFETY: A GUIDE FOR THE RESPONSIBLE DESIGN AND IMPLEMENTATION OF AI SYSTEMS IN THE PUBLIC SECTOR 63 (2019).

183. *See* ARTICLE 36 REVIEWS: DEALING WITH THE CHALLENGES*, supra* note 9, at 33.

VIII. EMERGING POLICY GUIDANCE

There is no international guidance on how to best conduct V&V and legal reviews of AI-driven systems. Yet, over the last years, governmental agencies and private companies have started to elaborate "normative guideposts" regarding safe and ethical development and use of AI systems. While they do not directly apply to V&V and legal reviews, it is interesting that they address the same or similar issues and challenges as identified and discussed above with regard to ensuring IHL compliant use of AI systems. As such, these emerging policies and related work can serve as inputs for the reflection on the development of future guidance for conducting V&V, including legal assessment, of military AI systems.

As of September 2019, 84 documents containing ethical principles or guidelines for AI have been issued worldwide by the public and private sector.[184] More recently, the European Union Commission has adopted its own guidance on AI.[185] In October 2019, the U.S. DoD adopted the first policy guidance directly addressing military AI. The "AI Principles: Recommendations on the Ethical Use of Artificial Intelligence" establishes the five principles of responsibility, equitability, traceability, reliability, and governability.[186] Interestingly, these principles have a lot of commonalities with other emerging guidance. Indeed, there are certain overarching principles among the guiding documents, namely transparency, justice and fairness, non-maleficence, responsibility, and privacy.[187]

Transparency refers to the possibility of describing, inspecting, and reproducing the mechanisms through which decisions are made, how learning and adaptation to the environment occur and how data is governed.[188] This requires that the system has specific technical qualities for it to be defined as transparent, such as the ability to give reasons for its actions

184. Anna Jobin, Marcello Ienca & Effy Vayena, *The Global Landscape of AI Ethics Guidelines*, 1 NATURE MACHINE INTELLIGENCE, Sept. 2019, at 389, 391.

185. *Commission White Paper on Artificial Intelligence: A European Approach to Excellence and Trust*, at 1, COM(2020) 65 final (Feb. 19, 2020), https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf.

186. DEFENSE INNOVATION BOARD, *supra* note 27.

187. Jobin, Ienca & Vayena, *supra* note 184, at 394.

188. Amina Adadi & Mohammed Berrada, *Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI)*, 6 IEEE *ACCESS* 52138, 52141 (2018), https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=8466590.

("explainability").[189] Transparency can help investigations into AI actions, as it would enable humans to retrace the decision steps and interactions with the environment that caused the result, understand what occurred, and make sure similar errors or violations would no longer happen. When such systems undertake targeting functions or contribute to such processes, transparency is an essential prerequisite from an IHL perspective. Notably, for obligations of conduct, such as the principles of proportionality or precaution, which require balancing between different values and the execution of feasible measures in given circumstances, transparency enables a reviewer to access the process that led to a specific output; allows the reviewer to assess it in light of those obligations; and, in instances of violations, enables the attribution of responsibility.[190] If defined as a mandatory parameter to fulfill, an AI system's level of transparency would need to be assessed during its V&V.

Another common trait of the guidelines is a focus on the reliability of the AI system, i.e., whether the system appropriately, safely, and robustly acts within its domain.[191] This implies verifying whether the system responds safely to unanticipated situations and does not evolve in ways that are inconsistent with the original expectations.[192] In particular, "safety" refers to "freedom from risk which is not tolerable,"[193] where the notion of freedom stands for a low probability of occurrence of non-tolerable consequences.[194] This directly links to the above discussions on acceptable levels of system predictability as a parameter to be assessed in the V&V process. Similarly,

189. A 2019 survey found transparency featured in seventy-three out of eighty-four documents providing AI guidance. *See* Jobin, Ienca & Vayena, *supra* note 184, at 391. Such notion of transparency is narrowly conceived. On the contrary, a broader understanding would also encompass transparency as to the decisions made by humans during the design, development and validation of the system. E-mail from Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe, to the authors (Sept. 28, 2020, 23:35 CST) (on file with the authors).

190. *See*, *e.g.*, Human Rights Watch, *The Toronto Declaration: Protecting the Right to Equality and Nondiscrimination in Machine Learning Systems*, HUMAN RIGHTS WATCH (July 3, 2018), https://www.hrw.org/news/2018/07/03/toronto-declaration-protecting-rights-equality-and-non-discrimination-machine.

191. DEFENSE INNOVATION BOARD, *supra* note 27, at 36.

192. MICROSOFT CORPORATION, *supra* note 111, at 65.

193. Joint Working Group of the ISO Committee on Consumer Policy & IEC Advisory Committee on Safety, *ISO/IEC Guide 51:2014, Safety aspects—Guidelines for Their Inclusion in Standards*, ISO (Apr. 2014), https://www.iso.org/standard/53940.html.

194. Simen Eldevik, *AI + Safety*, DNV GL (Aug. 28, 2018), https://ai-and-safety.dnvgl.com/.

"robustness" means that an AI system is capable of recognizing and behaving correctly when exposed to different scenarios compared to those in which it was trained.[195] Also, AI systems must be able to "adequately deal with errors or inconsistencies during all life cycle phases," and they should be resilient against attacks and attempts to manipulate data and algorithms.[196]

A further common feature in recent policies is the requirement that the system is not biased, i.e., that the system is fair or equitable.[197] The non-discriminatory nature of the system's outcomes may be verified during the certification stage. As discussed above, since the requirement is tightly linked to data, some guidelines advise to ensure during the training phase that the data fed into the system is non-biased.[198]

Guidelines also seek to ensure some degree of human control over AI systems. The U.S. DoD establishes that humans should remain responsible for the development, deployment, use, and outcomes of such a system.[199] Similarly, the EU white paper on AI lists human oversight as necessary for high-risk AI applications, suggesting this would help to prevent the system from causing adverse effects.[200] According to the white paper, human oversight can consist of (1) revision and validation of the AI output by a human before it becomes effective; (2) revision and validation of the outcome after it becomes effective; (3) real-time monitoring of the system while operating along with the ability to intervene and deactivate the system; and (4) imposing operational constraints on AI systems during the design phase.[201] How

195. Dario Amodei et al., *Concrete Problems in AI Safety* 1, 3 (July 25, 2016), https://arxiv.org/pdf/1606.06565.pdf. *See also Commission White Paper*, *supra* note 185, at 20 (stating that AI systems "must be technically robust and accurate in order to be trustworthy").

196. *Commission White Paper*, *supra* note 185, at 20.

197. *See, e.g.,* DEFENSE INNOVATION BOARD, *supra* note 27, at 31–33; MICROSOFT CORPORATION, *supra* note 111, at 59–62.

198. *Commission White Paper*, *supra* note 185, at 18–19.

199. Besides the Defense Innovation Board's recommendations on AI, DoD Directive 3000.09 on autonomy already provides "[a]utonomous and semi-autonomous weapon systems shall be designed to allow commanders and operators to exercise appropriate levels of human judgment over the use of force." *See* DoD Directive 3000.9, *supra* note 6, at 2.

200. *Commission White Paper*, *supra* note 185, at 21. According to the white paper, high risk AI applications are those deployed in high risk sectors, such as healthcare, transport, or energy; those deployed for high risk uses, such as AI applications that produce significant legal or injury risks; and those that produce effects that cannot be reasonably avoided. An analogy may be drawn between such high risk AI applications and those with similar features and implications in the military domain.

201. *Id.*

to configure human-machine interaction remains intensively debated.[202] Once the necessary level of human control is defined, it becomes a relevant parameter for testing, verifying, and validating AI systems. Depending on the defined degree and type of control, the assessment most likely would need to focus on the human-machine teaming rather than focusing on the system only. This would allow an ideal evaluation of the human-machine integration, including the humans' reliance and dependability on the system.

Although these principles are not legally binding and mostly still to be further developed at this stage, their operationalization may offer valuable indicators for experts and lawyers concerned with verifying AI systems for their proper use. Indeed, they may serve both to help achieve the above discussed technical, operational, and legal requirements for the use of military

---

202. *See*, *e.g.*, VINCENT BOULANIN ET AL., LIMITS ON AUTONOMY IN WEAPON SYSTEMS: IDENTIFYING PRACTICAL ELEMENTS OF HUMAN CONTROL 8 (2020), https://www.sipri.org/sites/default/files/2020-06/2006_limits_of_autonomy.pdf; United States, Human-Machine Interaction in the Development, Deployment and Use of Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, U.N. Doc. CCW/GGE.2/2018/WP.4 (Aug. 28, 2018), https://www.unog.ch/80256EDD006B8954/(httpAssets)/D1A2BA4B7B71D29FC12582F6004386EF/%24file/2018_GGE+LAWS_August_Working+Paper_US.pdf; Marc C. Canellas & Rachel A. Haga, *Toward Meaningful Human Control of Autonomous Weapons Systems through Function Allocation*, 2015 IEEE INTERNATIONAL SYMPOSIUM ON TECHNOLOGY AND SOCIETY (ISTAS) 1 (2015); M.L. Cummings, *Man vs. Machine or Man + Machine?*, 29 IEEE INTELLIGENCE SYSTEMS 62 (2014). Already in 2016, at the Group of Government Experts on LAWS the U.S. delegation, concerning the notion of the appropriate level of human judgments, affirmed that:

> [T]there is no "one-size-fits-all" standard for the correct level of human judgment to be exercised over the use of force with autonomous and semi-autonomous weapon systems . . . . In particular, the level of human judgment over the use of force that is appropriate will vary depending on factors, including the type of functions performed by the weapon system; the interaction between the operator and the weapon system, including the weapon's control measures; particular aspects of the weapon system's operating environment (for example, accounting for the proximity of civilians), the expected fluidity of or changes to the weapon system's operational parameters, the type of risk incurred, and the weapon system's particular mission objective. In addition, engineers and scientists will continue to develop technological innovations, which also counsels for a flexible policy standard that allows for an assessment of the appropriate level of human judgment for specific new technologies.

Michael Meier, U.S. Delegation Statement on "Appropriate Levels of Human Judgment" at the Informal Meeting of Experts on Lethal Autonomous Weapons Systems (Apr. 12, 2016), https://geneva.usmission.gov/2016/04/12/u-s-delegation-statement-on-appropriate-levels-of-human-judgment/.

AI systems, as well as offer additional elements States may wish to consider. For instance, many private initiatives, such as the IEEE,[203] are currently working on how to translate them into technical standards, which would allow safety and ethical principles to be effectively reflected in an AI system's functioning. Indeed, each of the principles can be linked to some methodologies, techniques, or more precise standards, which may guide an AI system's certification in terms of techniques to adopt and parameters to assess. It is noteworthy, however, that these principles currently remain aspirational as it is practically impossible to fulfill them all. A tradeoff between some of these principles also exists, whereby implementing one may imply renouncing (or, at least, partially renouncing) the implementation of another.[204]

In any case, the military use of AI requires further efforts to develop guidance for effectively evaluating and authorizing their use. States can do this both nationally and through international cooperation. This is in line with calls for good practices by the United States and the Group of Governmental Experts on LAWS, among others.[205] As the private sector continues to be progressive in developing implementable guidance for civil applications for AI, defense agencies may take inspiration and borrow insights. This could pave the way for a common understanding of what is required of AI systems performing military functions or supporting armed forces.

IX. CONCLUSION

With the emergence of military AI technology, the international legal framework needs to be digitalized. Regarding AI-enabled weapon systems, further research is required on how AI can effectively operationalize IHL. Yet the technology brings more than this to the legal regime on weapons. Rather than being just a military tool to be assessed according to Article 36 API and the respective customary international law rule, AI's features are such that targeting law needs to be integrated, and the system needs training in IHL based on data. The IHL rules to be verified for compliance by the weapon system are thus expanded during the legal assessment. This assessment must be done at an earlier stage in comparison to traditional weapon systems.

---

203. *See, e.g.,* notably, IEEE ETHICS IN ACTION IN AUTONOMOUS AND INTELLIGENT SYSTEMS, IEEE, https://ethicsinaction.ieee.org/ (last visited Feb. 22, 2021).

204. Telephone Interview with Ricardo Chavarriaga, Head, Office of the Confederation of the Laboratories for Artificial Intelligence Research in Europe (Nov. 11, 2020).

205. *See supra* notes 55, 57 and accompanying text.

Yet, the characteristics of AI technology fundamentally alter the legal assessment. As this technology is digital, the legal parameters need to be integrated and absorbed by the algorithm. Consequently, the legal review needs to assess its technical functioning as an algorithm. The evaluation of whether the system operates in accordance with its technical specifications can thus include the conformity with legal requirements. Hence, the legal review can be integrated into the V&V process of the AI system. The legal review thereby conflates with the technical process of V&V.

The comprehensive nature of such an assessment of AI systems leads to a two-fold consequence. First, legal advisors and technical experts need to cooperate during the testing and verification procedure of AI systems to ensure that the system correctly reflects its specifications and thus operates in accordance with IHL. For machine learning techniques, this includes the participation of legal advisors during data selection. Second, it implies that the predictability problem needs to be solved at the technical and operational level. This requires predetermining levels of predictability for each system and operational environment that indicate the acceptable level of risk that a given system might not comply with the applicable rules. This would reflect a new parameter against which to assess an AI system as a precondition for respecting IHL.

The law on legal reviews of weapons, in particular Article 36 API and the applicable customary international law norm, does not need to change. Article 36 API's formulation allows the adaption of State practice to the characteristics and challenges of the new technologies. As the rapid development of AI systems and lawyers' understanding of these technologies only began recently, further research, debate, and practical guidelines regarding the legal review of algorithms are necessary. Standards will need to be developed to define actionable directives on how to verify such systems for their safety and legality. Emerging guidelines regarding the development and use of AI, as well as respective initiatives from the private sector can inform such work.

The most significant implication of the emerging AI technologies regarding legal reviews is that they increase the overall importance of such reviews as mechanisms to ensure compliance with IHL. The more humans delegate crucial tasks to autonomous systems, the more the V&V and legal reviews become the essential gatekeeper for IHL compliance. The burden falls not only on legal advisers but also technical experts, as only together can they ensure that AI systems are legally operationalizable. Indeed, it is crucial that

legal reviews are rigorously conducted if they are to fulfill their promise of preventing IHL violations by new technologies.