SCIENCE & DIPLOMACY

Published on *Science & Diplomacy* (https://www.sciencediplomacy.org)

New Article

The Weapon that Mistook a School Bus for an Ostrich

Vulnerabilities in Artificial Intelligence and Challenges for the Meaningful Human Control of Autonomous Weapons
By Daniele Amoroso, Denise Garcia, and Guglielmo Tamburrini - 05/05/2022

The major powers are busy incorporating AI technologies into existing and emerging military systems in an ongoing and fast-developing AI military race.[1] The push towards a comprehensive militarization of AI is epitomized by the 2021 call to integrate AI-enabled technologies into “every facet of warfighting” by the US National Security Commission on Artificial Intelligence.[2] Strikingly similar wording had already been employed in the China’s 2017 “New Generation Artificial Intelligence Development Plan,” which underscored the need to “promote all kinds of AI technology to become quickly embedded in the field of national defense innovation.”[3] Russian President Vladimir Putin has also unabashedly claimed that whoever becomes the leader in AI will rule the world.[4]

The race to weaponize AI was initially fueled by the rise of autonomous weapons systems (AWS). These are AI-enabled weapons systems that “select and apply force to targets without human intervention.”[5] Instances include “loitering munitions,” namely munitions that overfly an assigned area in search of targets to dive-bomb and destroy without requiring any further human intervention after their activation. The loitering munition Kalashnikov ZALA Aero KUB-BLA has allegedly been used by Russian forces in Ukraine.[6] And the Turkish unmanned aerial vehicle STM Kargu-2 was reportedly employed in autonomous attack mode during the Second Libyan Civil War against Haftar-affiliated forces.[7]

Ethically and legally troubling aspects of autonomous weapons systems have been extensively discussed in both scholarly and diplomatic debates for over a decade. A consensus has emerged within the international community of states that all weapons systems, including autonomous ones, should be subject to “meaningful human control.”[8] However, what such control amounts to operationally depends crucially on technological developments in the field of AI. Moreover, the intentional exploitation of AI systems’ vulnerabilities to perturb autonomous weapons’ behaviors was scarcely considered a serious possibility until recently. The maturing of techniques to launch so-called “adversarial attacks” to AI systems that are developed by means of machine learning methods is changing this picture and related issues of human control. Adversarial attacks generate deceptive inputs that are designed to cause mistakes in the predictions or classifications made by these AI systems.

Here we explore challenges that AI’s emerging vulnerabilities to adversarial attacks raise for the debate over norms for AWS and for “meaningful human control” over AI-powered weapons systems. Normative debates are lagging behind technological developments. The militarization of AI must be regulated by ensuring that emerging challenges to meaningful human control are properly addressed by the international community of states and that adequate and effective regulations are developed, promulgated, and enforced.

Adversarial attacks and humans as “fail-safe” actors

To begin with, the main ethically and legally troubling aspects of autonomous weapons’ operational deployment are as follows:

i. Machines may violate the principles of distinction and proportionality enshrined in international humanitarian law, that is, the law that sets limits on permissible behaviors in warfare. The principle of distinction requires belligerent parties to always distinguish between combatants and non-combatants, and to direct attacks only against combatants. The principle of proportionality prohibits to launch attacks that are expected to cause incidental losses of civilian life, injuries to civilians or damage to civilian objects which are excessive in relation to the anticipated military advantage.
ii. There is a gray area regarding the moral and legal responsibilities by military commanders and operators for the use of autonomous weapons, including accountability gaps arising from the distinction and proportionality principles described above.
iii. A machine making life-or-death decisions is an affront to the human dignity of their victims. Within the international community, the very idea of machines making life-or-death decisions is largely regarded as repugnant, raising concerns about the compliance of lethal autonomous weapons systems with the intimation to protect populations and belligerents in accordance with the dictates of public conscience set forth, in international law, by the principle known as the “Martens Clause.”[9]

These concerns have been raised in connection with both existing weapons systems, such as the loitering munitions described above, and forthcoming ones, like swarms of small, low-cost autonomous drones that are still under development. In contrast, autonomous systems that are purely anti-materiel and defensive, like the Israeli Iron Dome or the US Phalanx, are not considered problematic; their operational autonomy is in fact favorably viewed from both military and humanitarian perspectives, since, by reacting at faster-than-human speeds, they can protect inhabited areas, vehicles, and buildings.

Discussions of the ethical and legal dimensions of autonomous weapons have increasingly focused on the distinctive role that the “human element” must play in the use of force,[10] and notably the idea that any weapons system—including AWS—must be subject to meaningful human control.

Clearly, to exert meaningful human control over AI-powered weapons systems, sources of perturbation that may disrupt AI systems behaviors must be addressed.[11] Consider those AI systems that are based on deep neural network architecture, which play a central role in current AI technological development and applications.[12] Adversarial experiments have unveiled special fragilities of these systems. For instance, after slightly modifying input images of school buses and turtles, systems of this kind were found to mistake school buses for ostriches[13] and turtles for rifles.[14] These misclassifications might have catastrophic consequences in warfare, insofar as normal uses of school buses are protected by the principle of distinction under international humanitarian law. Moreover, a human operator would never mistake a bus for an ostrich in the presence of those small variations and adjustments of inputs that adversarial methods exploit to fool AI models. Thus, the proper exercise of meaningful human control may prevent the occurrence of these disastrous outcomes.

Adversarial AI attacks are now being systematically carried out against AI systems operating in the real world. Notably, by altering the illumination of a stop signal on the street in ways hardly perceptible to the human eye, an AI system was induced to read it as a 30-mph speed limit sign.[15] To carry out this optical attack, AI scientists used only a low-cost projector, camera, and computer. In other words, inexpensive and readily available equipment is sufficient to launch adversarial AI attacks. These technological developments challenge the US's International Traffic in Arms Regulation and similar export control regimes. More important for our present concerns, these developments pave the way to intentional adversarial attacks on the battlefield, inducing AI-powered autonomous weapons to make perceptual mistakes by exploiting their vulnerability to small and difficult-to-detect input variations. Manipulation of visual objects and other inputs perceived by autonomous weapons systems might induce friendly fire lead to violations of international humanitarian law.

Similar hostile motivations may prompt intentional attacks of a different kind on autonomous weapons systems, carried out by "poisoning" their AI learning modules which corrupt datasets for learning, degrading the learning algorithm or the resulting AI model. There are no patches available to avoid either input manipulation or poisoning attacks, insofar as these are based on inherent weaknesses of the deep learning methods and systems that are prevalent today.[16] These manipulation risks are exacerbated by emerging model inversion attacks.[17] These sorts of attacks probe an already trained AI system and extract information about its training data, which might then be exploited to fool the trained system.

It has been suggested that properly functioning autonomous weapons will, in some distant future, come to match or even statistically surpass the performance of competent and conscientious human combatants.[18] This speculation does not exclude the possibility that even more sophisticated AI-powered weapons may commit ethically and legally disastrous errors due to the brittleness of AI systems to in responding to changing contextual situations that may fall outside the scope of a narrow set of assumptions and boundary conditions.[19] Such errors might be avoided by substantively involving a human operator in the decision-making loop. For this reason, human control has been properly invoked as a "fail-safe" mechanism.[20] The possibility for autonomous weapons without this form of human control to be beneficial is doubtful on various other grounds as well. For instance, the urban environments and densely populated centers where conflicts increasingly take place today make it complicated for algorithms to fully abide by the vast array of rules and norms set by international humanitarian law that limits the conduct of warfare. All of this compels a serious evaluation of the calls to reduce the role of human decision-makers in warfare.

Interpretability and explainability issues

How data gets represented and how information is processed in sophisticated AI systems are mostly opaque to human users.[21] The black-box nature of these systems poses special challenges to the exercise of meaningful human control over weapons systems. Indeed, if humans are not expected to blindly trust the machine, they should have a sufficient level of humanly understandable information about what data the machine is processing (interpretability requirement).[22] To achieve adequate situational awareness, they should additionally obtain an account of why the machine is suggesting or undertaking certain courses of action (explainability requirement). However, to develop interpretable and explainable AI systems powered by machine learning methods in general, and by deep learning methods in particular — is a formidable research problem, which now characterizes the overall goals of the explainable AI (or XAI in brief) research area.

Pending significant breakthroughs in XAI, one cannot but acknowledge the difficulty of ensuring the levels of system explainability that are *necessary* to achieve the situational awareness required to exert meaningful human control over AI-powered weapons systems. And such efforts are further complicated by the development of *adversarial* XAI.

Adversarial XAI techniques disrupt explanation capabilities of AI learning systems, by perturbing their inputs in ways that are not perceptible to humans, and yet radically altering the explanations provided by the machine. For example, in one study, the image of a truck on the road was slightly manipulated and fed into an AI image classifier endowed with an explanation module. The system classified correctly the manipulated input image as the image of a truck. However, the manipulated input induced the system to explain the provided correct classification solely in terms of the cloudy sky in the background. No salient features of the truck appearing in the foreground were included in this explanation.[23] Clearly, unreliable explanations of the targeting selection and engagement processes by autonomous weapons jeopardize the situational awareness needed to exert meaningful human control. Should adversarial attacks on the explanation capabilities of AI systems be exploited in relation to autonomous weapons, the competition between the development of more interpretable AI systems and their disruption may eventually feed a new spiral in the AI military race. Indeed, AI experts may engage into adversarial activities to identify and remedy weaknesses of the explanation modules of autonomous weapons. Adversaries may respond by developing more powerful adversarial techniques challenging this next generation of autonomous systems, and so on. In the end, these developments are paving the way to a multifaceted AI military race and loom large in the problem of meaningful human control.

The AI militarization race and the regulation of autonomous weapons systems

The exploitable weaknesses of autonomous weapons raise new questions and challenges for understanding the operational content of meaningful human control and its applicability. To illustrate, let us consider the official position of the International Committee of the Red Cross (ICRC), which functions as the guardian of international humanitarian law.[24] The ICRC has been a constant actor in the diplomatic talks on regulating autonomous weapons which began in Geneva in 2013 and, in 2021, published its official stance. This stance carries weight in influencing the positions of states and galvanizing the majority of member states to call for a new legally binding treaty.

The ICRC position is characterized by a *differentiated* approach to regulating autonomous weapons, with its selective call for the prohibition of certain types of weapons and the restriction and regulation of others. To begin with, they contend that anti-personnel autonomous weapons systems should be prohibited. This ICRC standpoint is motivated by a fundamental ethical concern about machines making life-and-death decisions and by concerns regarding compliance with the laws of war. Recasting this prohibition in terms of human control requires that human operators must always be in the position to decide what to do when an operation hinges on the question of whether one or more human beings should be targeted, and why. Thus, lethal weapons systems cannot operate in autonomous mode but rather as decision support systems. It bears noting that AI-powered classification and prediction capabilities of these decision support systems can still be attacked, producing disinformation or incorrect explanations for suggested choices. The challenge, therefore, for meaningful human control is to clarify what it means to take every possible precaution to avoid the threats posed by such attacks.

A similar challenge arises in connection with the additional ICRC recommendation to prohibit unpredictable autonomous weapons systems, namely those that are "designed or used in a manner such that their effects cannot be sufficiently understood, predicted and explained."[25] The operational implications of this general line drawn by ICRC may change on account of emerging possibilities for adversarial attacks, in addition to considerations concerning battlefield conditions of use and other known limitations of AI systems in safety-critical domains.[26] In particular, AI systems that are otherwise understandable, predictable, and explainable may abruptly lose these properties due to malicious attacks.

In connection with other autonomous weapons systems, the ICRC recommends that their use should be regulated through a combination of limits on the types of targets allowed; parameters for the duration, geographical scope, and scale of use; restrictions on situations of use, and "requirements for human–machine interaction, notably to ensure effective human supervision, and timely intervention and deactivation."[27] To all of this must be added the requirement that best efforts are taken to ensure protection from attacks exploiting AI inherent vulnerabilities that cannot be readily repaired. Clearly, speed is paramount in invoking and applying such efforts and setting norms to keep pace with rapid technological developments.

Finally, another major normative challenge is to devise mechanisms for transparency and verification that can ensure compliance with the requirement of human control, once such a requirement is internationally agreed upon. Unlike existing arms control treaties, a legal instrument on autonomous weapons systems would have to address the "process" by which a weapon is used, with a particular focus on the human-machine interaction.[28] This raises unprecedented problems regarding verification, since human control is "a qualitative feature, the human role in the target selection and engagement is not visible from the outside, and the software might be altered after inspection."[29] A promising way to overcome these difficulties is represented by the "glass box" approach suggested by the International Committee for Robot Arms Control (ICRAC). Under this approach, a treaty on autonomous weapons systems should include provisions obliging state parties to design weapons systems so that they automatically record and securely store, for each engagement, information that is relevant to demonstrate that operators were involved in compliance with human control requirements. Such information should be then made available to a "Treaty Implementing Organization, on request, when sufficient evidence exists to support suspicions of illegal autonomous operation."[30]

Conclusion

In February 2019, the Defense Advanced Research Projects Agency (DARPA) launched the GARD (Guaranteeing AI Robustness against Deception) program with a view to developing "a new generation of defenses against deception attacks on machine learning."[31] The announcement, however, did not include any references to adversarial XAI, since, at that time, studies on the matter were still in their infancy.

This absence highlights the formidable conceptual and operational challenges in determining what is needed to apply meaningful human control to weapons systems, challenges that are amplified and transformed by an ongoing AI militarization race geared towards identifying and exploiting the vulnerabilities in most AI-based systems. This militarization race shows how pressing is to move forward in diplomatic negotiations about the regulation of AWS. However, actions of the international community of states lag alarmingly behind the technological development of AWS and their enabling technologies. For instance, at the 6th Review Conference of the State Parties to the Convention on Certain Conventional Weapons (CCW), held December 13–17, 2021, at the United Nations in Geneva, state parties to the CCW limited themselves to adjourning the work of the Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems established in 2016, providing neither a clear mandate for the GGE to negotiate a multilateral instrument on the issue, nor the slightest indication about its prospective contents.

It is vitally important to anticipate, recognize, and fully account for the changing landscape of the human control problem. This refers to the unintended consequences of unregulated automation at a grand scale, and the insidious march towards a situation where humans will have relinquished oversight. Proactive actions must be taken to ensure that issues that emerge are properly and adequately addressed and that effective regulations of autonomous weapons are developed, promulgated, and enforced. Such regulation should be aligned with the goals of international bodies such as the UN Secretary-General, the ICRC, ICRAC and other NGOs involved in the worldwide Campaign to Stop Killer Robots, as well as a majority of the United Nations member states. Moreover, scientific and theoretical uncertainties surrounding proven and potential AI vulnerabilities militate in favor of a precautionary approach to the regulation of autonomous weapons systems.[32] This cautious approach sets a high bar of human control in order to reduce the likelihood that emergent or newly discovered AI vulnerabilities could be exploited and to keep intact the "fail-safe" role of human decision-makers.[33]

Endnotes

1. Heather M. Roff, "The Frame Problem: The AI "Arms Race" Isn't One," *Bulletin of the Atomic Scientists* 75, no. 3 (2019): 95–98; Denise Garcia, "Stop the Emerging AI Cold War," *Nature* 593, no. 7858 (2021): 169.
2. US National Security Commission on Artificial Intelligence (NSCAI), "Final Report," 2021, www.nscai.gov/wp-content/uploads/2021/03/Full-Report-Digital-1.pdf.
3. China's State Council (2017), "New Generation Artificial Intelligence Development Plan" (translation), www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translat....
4. *Russia Today*, "Whoever Leads in AI Will Rule the World": Putin to Russian Children on Knowledge Day," September 1, 2017, www.rt.com/news/401731-ai-rule-world-putin.
5. US Department of Defense, "Autonomy in Weapons Systems (Directive 3000.09)," November 21, 2021, www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf; International Committee of the Red Cross, "ICRC Position on Autonomous Weapons Systems," May 12, 2021, www.icrc.org/en/document/icrc-position-autonomous-weapon-systems.
6. Zachary Kallenborn, "Russia May Have Used a Killer Robot in Ukraine. Now What?" *Bulletin of the Atomic Scientists*, March 15, 2022, https://thebulletin.org/2022/03/russia-may-have-used-a-killer-robot-in-ukraine-now-what.
7. United Nations, "Final Report of the Panel of Experts on Libya Established Pursuant to Security Council Resolution 1973 (2011)," March 8, 2021, UN Doc. S/2021/229, para. 63, https://digitallibrary.un.org/record/3905159?ln=en. The list of existing types of autonomous weapons is continually expanding, with an initial comprehensive survey provided in Vincent Boulanin and Maaike

Verbruggen, *Mapping the Development of Autonomy in Weapon Systems* (Solna: Stockholm International Peace Research Institute, 2017).

8. Ray Acheson, "Editorial: Convergence Against Killer Robots," *CCW Report* 9, no. 3 (August 8, 2021); Frank Sauer, "Lethal Autonomous Weapons Systems," in *The Routledge Handbook Social Science Handbook of AI*, ed. Anthony Elliott (London: Routledge, 2021), 237–250.
9. The Martens Clause requires that, in the absence of applicable treaty provisions, "populations and belligerents remain under protection and empire of the principles of international law, as they result from the usages established between civilized nations, from the laws of humanity, and the requirements of public conscience." For discussion, see Daniele Amoroso, *Autonomous Weapons Systems and International Law: A Study on Human-Machine Interactions in Ethically and Legally Sensitive Domains* (Baden-Baden: Nomos Verlagsgesellschaft, 2020).
10. International Panel on the Regulation of Autonomous Weapons (iPRAW), "Building Blocks for a Regulation on LAWS and Human Control: Updated Recommendations to the GGE on LAWS," July 2021, www.ipraw.org/wp-content/uploads/2021/07/iPRAW-Report_Building-Blocks_July2021.pdf.
11. Mary L. Cummings, "Rethinking the Maturity of Artificial Intelligence in Safety-Critical Settings," *AI Magazine* 42, no. 1 (2021): 6–15.
12. Gary Marcus, "Deep Learning: A Critical Appraisal," arXiv.org, last updated January 2, 2018, https://arxiv.org/abs/1801.00631.
13. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus, "Intriguing Properties of Neural Networks," arXiv.org, last updated February 19, 2014, https://arxiv.org/abs/1312.6199.
14. Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok, "Synthetizing Robust Adversarial Examples," arXiv.org, last updated June 7, 2018, https://arxiv.org/abs/1707.07397.
15. Abirham Gnanasambandam, Alex M. Sherman, and Stanley H. Chan "Optical Adversarial Attack," IEEE/CVF International Conference on Computer Vision Workshops (ICCVW), 2021, https://openaccess.thecvf.com/content/ICCV2021W/AROW/papers/Gnanasambandam_Optical_Adversarial_Attack_ICCVW_2021_p
16. Marcus Comiter (2019), "Attacking Artificial Intelligence: AI's Security Vulnerability and What Policymakers Can Do About It," Belfer Center for Science and International Affairs, Harvard Kennedy School, August 2019, www.belfercenter.org/publication/AttackingAI.
17. Qian Wang and Daniel Kurz, "Reconstructing Training Data from Diverse ML Models by Ensemble Inversion," arXiv.org, November 5, 2021, https://arxiv.org/abs/2111.03702. For a more comprehensive list and analysis of AI adversarial techniques, see Muhammad Mudassar Yamin, Mohib Ullah, Habib Ullah, Basel Katt, "Weaponized AI for Cyberattacks," *Journal of Information Security and Applications* 57, (2021): 102722.
18. Ronald C. Arkin, *Governing Lethal Behavior in Autonomous Robots* (Boca Raton, FL: CRC Press, 2009).
19. Cummings, 2021.
20. Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W.W. Norton, 2018); Daniele Amoroso and Guglielmo Tamburrini, "Autonomous Weapons Systems and Meaningful Human Control: Ethical and Legal Issues," *Current Robotics Reports* 1, no. 4 (2020): 187–194.
21. Arthur Holland Michel, "The Black Box, Unlocked: Predictability and Understandability in Military AI," United Nations Institute for Disarmament Research, 2020, https://unidir.org/publication/black-box-unlocked.
22. Patrick Chisan Hew, "Preserving a Combat Commander's Moral Agency: The Vincennes Incident as a Chinese Room," *Ethics and Information Technology* 18, no. 3 (2016): 227–235.
23. Amirata Ghorbani, Abubakar Abid, and James Zou, "Interpretation of Neural Networks is Fragile," *Proceedings of the AAAI Conference on Artificial Intelligence* 33, no. 1 (2019): 3681–3688.
24. ICRC, 2021.
25. Ibid.
26. Cummings, 2021.
27. ICRC, 2021.
28. International Panel on the Regulation of Autonomous Weapons (iPRAW), "Verifying LAWS Regulation: Opportunities and Challenges," August 2019, 3, www.ipraw.org/wp-content/uploads/2019/08/2019-08-16_iPRAW_Verification.pdf.
29. Ibid.
30. International Committee for Robot Arms Control (ICRAC), "Compliance Measures for an Autonomous Weapons Convention," ICRAC Working Paper #2, May 2013.
31. https://www.darpa.mil/program/guaranteeing-ai-robustness-against-deception.
32. Denise Garcia, "Future Arms, Technologies, and International Law: Preventive Security Governance," *European Journal of International Security* 1, no. 1 (2016): 94–111.
33. Daniele Amoroso and Guglielmo Tamburrini, "Toward a Normative Model of Meaningful Human Control Over Weapons Systems," *Ethics & International Affairs* 35, no. 2 (2021): 245–272.

Tags: emerging technologies, autonomous weapons, AI

Source URL: https://www.sciencediplomacy.org/article/2022/weapon-mistook-school-bus-for-ostrich?utm_campaign=PANTHEON_STRIPPED&utm_source=PANTHEON_STRIPPED&utm_medium=PANTHEON_STRIPPED