Webinar series on emerging technologies
in the area of lethal autonomous weapons systems
Technological Aspects – 26 October 2020

In October 2020, the United Nations Institute for Disarmament Research and United Nations Office for Disarmament Affairs convened three webinars to inform the ongoing deliberations of the group of governmental experts on emerging technologies in the area of lethal autonomous weapons systems (the GGE on LAWS). The webinars considered relevant technological, military, and legal aspects that would benefit from additional clarification or review.

The first webinar addressed technological aspects of LAWS. It was convened virtually by the United Nations Institute for Disarmament Research on 26 October 2020.

Summary of the webinar on technological aspects

Background

In November 2019, the Group of Governmental Experts on emerging technologies in the area of Lethal Autonomous Weapon Systems (GGE on LAWS) received a mandate for 2020 and 2021 to produce "*consensus recommendations in relation to the clarification, consideration and development of aspects of the normative and operational framework on emerging technologies in the area of lethal autonomous weapons systems*"[1].

A key enabler for implementing this mandate is a more mature understanding of the technologies that would be the subject of any such normative and operational framework. To facilitate a more mature common technical understanding and support further dialogue, this first webinar, with the help of a group of subject matter experts from academia, industry and government,[2] unpacked two elements of LAWS: the algorithmic core of AI and the issue of integrating this core component into physical systems.

Inside AI: understanding algorithms

Broadly speaking, algorithms are the computational elements that enable "autonomy" in a LAWS, in the sense that they are responsible for an autonomous system's capacity to process and act upon inputs from the environment. In general terms, algorithms can be defined as a process or set of rules to be followed in calculations or other problem-solving operations, including the performance of a specific task. However, it is important to note that AI systems that have driven many of the most recent advances in autonomy, and in particular in those based on Machine Learning, algorithms operate at a different, less deterministic, level: those who build the algorithms do not

[1] Final Report of the Meeting of the High Contracting Parties to the Convention on Prohibitions or Restrictions on the Use of Certain Conventional Weapons Which May Be Deemed to Be Excessively Injurious or to Have Indiscriminate Effects. CCW/MSP/2019/9. 13 December 2019.

[2] Experts included: Dr. Vanina Martinez (University of Buenos Aires), Dr. Ricardo Rodriguez (University of Buenos Aires), Jane Pinelis (Joint Artificial Intelligence Center, US), Barry O'Sullivan (University College Cork), Thomas Grohs (Airbus), Dr Gabriele Rizzo (FUTURES Lab) and Abhishek Gupta (Microsoft).

provide specific rules for how the system can accomplish a task, but rather they establish a framework for the system to build its own task-level models to solve specific problems.

At their current levels of technological maturity, these kinds of AI perform well in basic cognitive tasks (e.g. computer vision, voice recognition and natural language processing). However, an important difference should be noted between AI's ability to process large amounts of data and its ability to understand such data. Today's AI are not yet able to "understand" and contextualize the data they process, nor can they necessarily assess causal relationships between observed events or features. Therefore, they are better at executing narrowly defined tasks (e.g. pattern recognition).

The panelists also noted that improvements in AI performance in one specific task or application are not automatically transferable to other tasks or applications. For example, the fact that over the past years AI has been able to defeat humans in increasingly complex games does not mean that AI will therefore be able to perform well in other applications, especially if such applications occur in an operational environment in the physical world. As a consequence, it is difficult to imagine a near future in which AI can fully substitute human decision-makers in any complex judgment or reasoning tasks.

That being said, AI remains a rapidly expanding field of active research and development. As more researchers and developers build upon each others' work and seek to further refine and improve algorithm performance, we will develop a much more thorough understanding of the technology's strengths and limitations across the range of applications for which we imagine it being useful. It is generally agreed that future advances in the technical field could unlock new areas of progress and enable more disruptive applications.

With respect to learning-based AI, given that such systems are trained rather than programmed, the data used to train such algorithms play a fundamental role in determining the performance and behavior of the system: AI is only as good as the teacher (i.e. training data). When an autonomous system encounters an input for which it has not been trained it is likely to fail; this characteristic of AI systems is known as "brittleness." Broadly speaking, such inputs arise from either environmental conditions that were not anticipated or errors induced by adversarial action (e.g. data poisoning or spoofing). As it is difficult, if not impossible, to guarantee that any given training dataset comprehensively and unbiasedly represents all the possible situations or adversarial actions that could potentially arise in the environment to which an autonomous system will be deployed, it is difficult to even conceptualize any AI application that is 100% predictable and reliable.

AI beyond Machine Learning

While the terms AI and Machine Learning (ML) are often used interchangeably, the latter is only a subset of the former—which includes a wide range of approaches that do not involve any form of algorithmic "learning". There are at least 40 years of work in AI that focused on different models that are not based on ML (e.g. argumentation).

The AI systems that will be employed in autonomous systems are likely to be "hybrid" systems that combine rule-based and non-deterministic (i.e. learning-based) algorithms. For example,

one such approach is known as Neuro Symbolic (NS) AI, which "is aimed at augmenting (and retaining) the strengths of statistical AI (machine learning) with the complementary capabilities of symbolic or classical AI (knowledge and reasoning)", with a view "to solve much harder problems, learn with dramatically less data and for a large number of tasks, and provide for inherently understandable and controllable decisions and actions." [3]

Fueling AI for training and operational use: the critical importance of data

AI systems depend upon models that are built, developed, trained, tested and evaluated with data. Despite the critical importance of data in both the development and operation of AI systems, including those systems that would feature in autonomous weapons, the issue of data has not so far received enough attention in the GGE on LAWS.

Discussion about AI systems that exclude considerations such as data availability, collection, curation, sharing, governance, storage and security are unlikely to adequately address the full scope of relevant policy issues. For example, if no data are available to train and test a system for a specific task, that system cannot ultimately be used to perform that task. Similarly, issues related to data are one of the main reasons why certain AI systems often perform well in the controlled environment of a lab, only to fail systematically when deployed in the real world: training or testing data does not exactly match the real world data (often referred to as data shift).[4]

In this context, panelists highlighted that traceability of access to data (i.e. who has access to data, when and under what circumstances) is an important ethical principle for AI. In particular, given the importance of testing and evaluation, testing data should, in principle, not be made available to vendors to avoid encouraging them to develop an application specifically designed to perform well in tests, but poorly in the real world.

Building on an observation from the virtual floor, experts discussed the merits of developing a shared international dataset specifically designed to test and evaluate AI applications with respect to international law (IL) and international humanitarian law (IHL). In principle, it was recognized that having a common dataset and baseline would be important not only for verification, validation, testing and evaluation, but also to allow meaningful exchanges of information between states. A counter argument for such a shared dataset would be related to the risks of sharing it with vendors for the reasons outlined in the preceding paragraph. In any case, a common dataset should be intended as establishing a baseline and would not substitute national testing datasets that are likely to remain confidential. While some practical challenges to implementation would require further thought, the idea of a common dataset was positively received by the experts, with the addition that, for technology developers, a complementary product such as a common ethical requirement

[3] See: https://researcher.watson.ibm.com/researcher/view_group.php?id=10518

[4] Recently, one research group proposed that another key reason for the gap in performance of ML systems between lab and real world may be a phenomenon known as "underspecification", which refers to the current inaccuracy of testing in identifying among the many models produced during the training of an AI system which ones will work better in the real world. For more information on underspecification please see: https://arxiv.org/pdf/2011.03395.pdf

document (e.g. 'handbook') would be useful to ensure compliance with legal and ethical principles by design.

The challenges of integrating AI in autonomous (weapon) systems

A whole new level of complexity is introduced in the discussion when we transition from considering AI in isolation to AI integrated with the many other elements that would constitute an autonomous (weapon) system. These elements include the sensors and communications devices that collect the data to be processed by the internal algorithms, other algorithms responsible for other autonomous functions, and the actuators that physically execute the "decisions" made by the AI component.

Addressing autonomy in this manner helps to shed light on some of the other core technical aspects considered in the GGE discussions, namely: the technical components necessary to enable humans to manage a system's autonomy through human-machine-interaction; the challenges of achieving appropriate levels of reliability and predictability when coupling complex algorithms within complex systems-of-systems, especially in light of data issues such as bias and spoofing; and the requirements for, and challenges of, realistic and sufficient testing and validation to mitigate risks.

Reality vs Fiction

Given the technical complexity of the subject, distinguishing between what is science and what, today, remains fiction can be quite challenging. During the webinar, experts were asked to articulate their view of what is reality and what remains fiction. While not an exhaustive list, the following applications provide at least an indication of the current technological maturity and expected/achievable developments on the horizon:

- *Reality:* complex data processing to support human decision making; human-level (or potentially even better) capabilities for specific, narrowly defined, tasks; reduced failure rates in certain applications or environments; replicability of specific features of human intelligence (accounting for the limitations imposed by system training); particularly relevant for non-state actors, the ability to build a system that brings together different AI sub-systems and integrate them in commercial-grade physical platform (e.g. 'stitching together' different AI systems in a sort of 'Frankenstein approach'.

- *Fiction:* complex interaction between crewed platforms and autonomous systems, particularly in domains, such as air, where platforms and operators are subject to very strict certification requirements; artificial 'general' intelligence capable of taking on any task with a general training and perform to a satisfactory level; embedding or programming 'consciousness' in AI (particularly important in the context of ethics); autonomous weapon systems that operate with humans off the loop.

It should be noted that the fact that an AI application may be scientifically feasible does not, by default, mean that such application would necessarily be operationally 'desirable'. There are a variety of reasons for this including, scalability, reliability, training requirements, cost factors, etc.

Exploring the role of the human

As previously mentioned, the role of humans remains key in both current and future applications of autonomous systems, particularly in the military domain. This is a reflection of the fact that the combination of humans and AI is more effective at some tasks or in some environments than AI or humans working individually. However, for humans to be able to work in tandem with autonomous systems effectively they have to be appropriately empowered and trained.

Efforts to achieve successful interaction between human operators and autonomous systems may be complicated by two opposite phenomena; on one side, 'automation bias' (i.e. excessive trust in a machine-generated output) and, on the other, 'algorithmic aversion' (i.e. excessive distrust in a machine-generated output). Hence the goal is to engender effective trust calibration: enabling the human to understand how to adjust their level of trust in the autonomous system based on a well-established understanding of the factors, circumstances and trade-offs that impact the system's performance (i.e. how much the operator should trust a system when it encounters certain kinds of inputs or behaves in certain ways).

Predictability and explainability are loose concepts used to refer to what could be called "trust". As mentioned above, 'calibrating trust' between users and machines is more accurate than 'building trust'. Such trust calibration is important to achieve the required predictability, which is important for warfighters who need confidence that certain inputs will result in certain outputs and understand which conditions and parameters will impact performance. As such, to operators/end-users, systems have to be understandable and intelligible more than 'explainable'.

A more detailed level of explainability might be more appropriate for the engineers testing the system earlier in the process. However, it should be noted that achieving detailed explainability might also be complicated by competing interests in protecting Intellectual Property rights owned by technology developers and vendors.

One of the panelists suggested that in human-machine interaction, automation and autonomy should be considered on different axes than responsibility and control, and that in no instances should the degree of autonomy of a system be considered as a substitute for human responsibility. However, for humans to retain appropriate levels of control, they need to be able to engage with the system efficiently and effectively, interpreting the signals that they receive: training and education are key. One expert raised the example of the fatal crash, in 2009, of an Air France flight: the black box showed that the airplane could have been saved from its stall but the human pilots took too long to understand the signals generated by the autopilot and were unable to intervene in time.

From lab to field: the importance of testing and evaluation

Testing and evaluation (along with verification and validation) are key for transitioning autonomous systems from development to actual real-world deployment in operations. The purpose of testing and evaluation (T&E) is not to build trust or ensure explainability: it is to

quantify risk. It is then up to decision makers and users to determine what level of risk is acceptable in a given circumstance, provided that they are equipped with the necessary knowledge to make such assessment in an informed way.

Methods for conducting T&E for autonomous systems equipped with AI components are radically different from the T&E process that is applied to traditional systems: the concept of having one final 'test event' to validate that technical requirements are met (e.g. range, speed, accuracy, endurance, etc.) which, if passed, results in the achievement of a certification is not applicable to systems featuring AI which require T&E to happen over time.

An iterative three-step approach, currently followed by the US Joint AI Center, was presented: the first step focuses on the AI component by itself, using test datasets to assess the accuracy of the AI algorithm. The second step is the system integration test which focuses on assessing the reliability of the system as a whole.[5] The third and last test is the human factors test that validates the degree to which successful trust calibration – along with other components of successful human-machine teaming – can be achieved.

Finally, it should be noted that the science behind T&E is still in development: AI technology develops faster than our ability to test it. More resources should be given to T&E to ensure this crucial step does not become a bottle-neck between the development and use of AI applications.

[5] A common metric to measure such reliability is Mean Time Between Failures (MTBF), which represents the predicted elapsed time between failures of a mechanical or electronic system during normal system operation.

SUGGESTED READINGS

During the webinar a number of resources were cited by the speakers as potentially of interest for member states and other stakeholders involved in the LAWS discussion:

Technical aspects of AI

- The Black Box, Unlocked https://www.unidir.org/publication/black-box-unlocked
- A Formal Framework for Agency and Autonomy https://www.aaai.org/Papers/ICMAS/1995/ICMAS95-034.pdf
- Self-driving Car https://en.wikipedia.org/wiki/Self-driving_car#Autonomous_vs._automated
- What is an Autonomous Car? https://www.synopsys.com/automotive/what-is-autonomous-car.html
- Does Object Recognition Work for Everyone? https://arxiv.org/pdf/1906.02659.pdf
- On Artificial Intelligence - A European approach to excellence and trust https://ec.europa.eu/info/sites/info/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf

- Robust Physical-World Attacks on Deep Learning Visual Classification https://arxiv.org/pdf/1707.08945.pdf
- Research summary: Sponge Examples: Energy-Latency Attacks on Neural Networks https://montrealethics.ai/research-summary-sponge-examples-energy-latency-attacks-on-neural-networks/
- Green Lighting ML: Confidentiality, Integrity, and Availability of Machine Learning Systems in Deployment https://arxiv.org/abs/2007.04693
- Learnability can be undecidable https://www.nature.com/articles/s42256-018-0002-3?error=cookies_not_supported&code=d829439a-e546-4e25-bde2-d809152eb63b

Ethics and AI

- The Moral Machine Experiment https://www.nature.com/articles/s41586-018-0637-6
- Moral Machine - Human Perspectives on Machine Ethics https://www.moralmachine.net
- IEEE Ethically Aligned Design principles: https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead1e.pdf
- IEEE Ethically Aligned Design in Practice https://ethicsinaction.ieee.org
- IEEE Ethics in Action in Autonomous and Intelligent Systems https://ethicsinaction.ieee.org/p7000/
- IEEE P7000™: Model Process for Addressing Ethical Concerns During System Design https://standards.ieee.org/project/7000.html
- IEEE P7000 - Engineering Methodologies for Ethical Life-Cycle Concerns Working Group https://sagroups.ieee.org/7000/

- Disruptive Technologies in Military Affairs https://lngv.ws/f2a
- State of AI Ethics June 2020 Report https://bit.ly/stateofaiethics1
- Weekly AI Ethics newsletter https://aiethics.substack.com
- AI Ethics blog from the Montreal AI Ethics Institute https://montrealethics.ai/blog
- Research summary: Overcoming Barriers to Cross-Cultural Cooperation in AI Ethics and Governancehttps://ethicsinaction.ieee.org/p7000/
- US DOD – JAIC: Ethical Principles for Artificial Intelligence https://www.ai.mil/docs/Ethical_Principles_for_Artificial_Intelligence.pdf
- Definitions of Intent for AI Derived From Common Law https://easychair.org/publications/preprint_download/GfCZ