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

19 August 2020

English only

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems

Geneva, 21-25 September and 2-6 November 2020

LAWS and human control: Brazilian proposals for working definitions

LAWS

1. The weaponization of Artificial Intelligence (AI) – the so-called algorithmic warfare, notably in association with robotics, cyber warfare, drone, and missile technology – has given rise to artifacts of singular nature notwithstanding century-old efforts to regulate the conduct of hostilities and the means of war. Since AI warfare has produced unique weapons, the issues raised by them must be addressed distinctively *vis à vis* conventional artifacts.

2. Autonomous Weapons Systems (AWS) are different for a set of reasons:

a) Being "intelligent," they are capable of evolving on their own, due to the functions of self-learning and self-(re)programming. This being so, they are essentially unpredictable in the long run, for some parameters imbedded in their software may be overruled and "improved" by the systems themselves. Therefore, intelligent machines may bypass human instructions and find breaches in command and control, what makes necessary the establishment of limits at an early stage of their design and development;

b) They are more performant, and increasingly more lethal;

c) From the engineering point of view, AI inserts an upper layer of abstraction above the system's programming language; by doing so, it widens the "cognitive distance" between the decision to activate an AWS and the consequences of the attack; since AI further isolates the human operator from the "heat of the battle," the user's perception and decision-making will tend to be more abstract and detached from the intuitions and emotions that arise from close contact with enemies;

d) While the operator may receive better information on the conflict environment, certain critical functions will be outsourced to the machine during the attack procedures (tracking, targeting, locking, engaging);

e) The environment that informs the operator is mathematically modeled and may be subject to misunderstandings and malfunctions; human errors may thus be replaced by cyber misinterpretations of the environment or situational awareness, or by system biases.

3. Given the extraordinary complexity of the subject matter and the rapid pace of AI technology involved, there is still no consensus on the definition of AWS. Nevertheless, technical complexities should not hinder progress in the discussion of LAWS governance, which should be based upon the concept of human-machine interaction, particularly human control, in compliance with IHL. The "conceptual trap" may be proved counterproductive in the long run, for IHL enhancement with regard to LAWS is in the interest of collective security.

4. Thus, Brazil favours a workable, pragmatic definition of LAWS, that goes beyond the "technology-centric definitional approach". The concept proposed by ICRC-SIPRI[1], elegant in its simplicity, is of great usefulness in this regard, and should be adopted by the GGE:

> *"Autonomous weapon system is any weapon system that once activated can select and attack targets without human intervention."*

5. For a more comprehensive definition of AWS, Brazil proposes the following addition: "An intelligent weapon system with autonomous operation mode (i.e., without human input after activation) capable of recognizing patterns in combat environments, and of learning to operate and make decisions regarding the critical functions of target identification, tracking, locking-on and engaging based on uploaded databases, acquired experiences and its own calculations and conclusions."

Human-machine interaction and human control

6. To what extent can algorithms, syntax and semantics of the programming language of AWS comply with the principles of distinction, proportionality, precaution, prohibition of indiscriminate attacks, protection of combatants and civilians, and reduction of collateral damage in the absence of human control?

7. Who will be held accountable for the misuse or the eventual unintended result of the use of an AWS?

8. What levels of unpredictability – a key feature of IA and AWS – are acceptable to IHL?

9. The above questions put the objective notion of human control at the center of the discussion on human-machine interaction and accountability in the use of AWS. The cornerstone of the work of the GGE must be the concept of human control instead of subjective concepts like "human judgment" and "intent."

10. Human-machine interaction is the link between, on the one side, engineering, and operational system, and the other, the operator. The machine, extension of the human operator, responds to the user's consciousness, judgment, knowledge, professional training, and intent.

11. This interaction takes place in two spheres: software, including programming language and database matching; and hardware, including drones, robots, missiles, or vehicles. Both areas of interaction follow strict rules of engagement and command and control, linking the operator to his superiors in compliance with military protocols and legal rules.

12. Since AI adds an upper layer of abstraction on top of the programming language, as mentioned earlier, the programmer and the operator do not have full control over the behaviour of the machine; instead, they set goals and rules that are read by the "inference engine", allowing the machine to take its own decisions according to those parameters. Thus, autonomous systems reduce the controlling role of the programmer, and even less control is left to the operator. Human control will be increasingly challenged by the sophistication of AWS, adding higher levels of unpredictability to the behavior of intelligent warfare if limits are not put in the earlier stages of their lifecycle.

13. After activating the device ("fire and forget"), the operator may not be totally sure of the ultimate target, or of the time and location of the attack. Since the machine behavior may be different from the user intent, there must be some level of "human on the loop" control in order to achieve the desired result.

[1] SIPRI-ICRC. Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control. Available at: https://www.sipri.org/publications/2020/other-publications/limits-autonomy-weapon-systems-identifying-practical-elements-human-control-0

14. Moreover, AWS receive inputs from the environment, which also may be misinterpreted by the system or changed after the moment of activation of the system.

15. Although AWS may provide better situational awareness and tactical-operational efficiency, as well as a much more accurate and efficient response in compressed time-frames (e.g., against missiles or lasers), human control exerted by combatants is necessary to make accurate judgments in the conduct of hostilities in order to both achieve military purposes and to assure compliance with IHL. This includes the possibility of intervening to override the machine's action and terminate engagements, especially in the event of system failure.

Human control as the cornerstone

16. AWS changed the place of users from manual operators to supervisors of the machine's operations. Since intelligent machines are "logical," but not "reasonable," lacking common sense and abstract thinking, and since they reduce the controlling function of programmers and users, humans must retain the ability to supervise, intervene and deactivate attack procedures, for they possess cognitive, holistic and intuitive capabilities that AWS do not have: qualitative judgment, reasoning, and reflection about the consequences of specific attacks. Moreover, the role of human sensibility in decisions that cause loss of lives and the destruction of houses, buildings, and facilities, should not be overlooked. Those complex capabilities cannot be inserted into AI systems, but they are inherently present in the minds and the personal experience of commanders and combatants within the framework of war protocols, rules of engagement, chains of command and control and interpretation of IHL rules.

17. The concepts of "human judgment" and "human control" are not only compatible but necessarily interlinked. They are not mutually exclusive, for they refer to different levels of the human-machine interaction (or teaming): "human judgment" involves the doctrine of employment, while "human control" is the operation of the weapon itself. Since it is not the scope of this paper and of the GGE mandate to discuss military doctrine – the realm of "human judgment" –, the focus should be put on the operation of the AWS – thus on "human control".

18. The objective concept of "human control" refers to the human-machine interface (HMI) and the modes of operation of the weapon: Off, Stand-by, Manual, Semi-auto, and Auto.

19. On its part, the broader and subjective concept of "human judgment" refers mainly to the discernment ability of the individuals under the chain of command and control (commanders, supervisors, operators) related to the weapon deployment, taking into account the doctrine, the habilitation of the various modes of operation, rules of engagement, training, and combat contexts.

20. However, to ensure that machines execute the intent of commanders and operators in the use of force solely on the basis of human judgment is not sufficient. Accountability must be required in the case of the unintended result of the use of an AWS: for instance, a requirement for the insertion of the supervisor's password to go from Semi-auto to Full Autonomous mode of operation.

21. Given the nature of AWS, the machine behavior may cause "unintended engagements" different from the user "intent," informed by the operator's "judgment," in the absence of human control. Human control is thus the sole concept capable of assuring the responsible use of AI in weapons systems. Responsibility, accountability, and liability in the event of unlawful employment caused by intent, guilt, deceit, recklessness, negligence, or malpractice must be ensured.

22. In synthesis, lawful AWS operations must rely not on "human judgment" or "intent" – which are essentially subjective –, but on the objective concept of "human control" over the critical functions and supervision to correct autonomous decisions that produce collateral damage, override system failures or misinterpretations of the environment, target, timing and to achieve the desired outcome both in military and legal terms.

23. A responsible chain of command and control cannot outsource the compliance with IHL – distinction, proportionality, precaution – and the moral and legal implications of unlawful use of force to inanimate machines, regardless of their sophistication and intelligence. Moreover, it is essential to clarify the causal link between the agent's conduct and the violation. Good faith and adequate judgment disconnected with meaningful control may not be sufficient to assure compliance with IHL rules in the operation of intelligent machines. Deployment of AWS involves a degree of risk assessment and responsibility that cannot be free from accountability under international law and IHL.

24. The already cited ICRC-SIPRI report underlines that human control can be exercised in three ways: controls on the AWS parameters, on the environment, and on human-machine interaction. The report also examines the phases when requirements for human control may be operationalized or implemented: study, research, and development, procurement, deployment. The GGE could further elaborate on how those controls could be translated into IHL parameters.

25. The discussion on human control should take into account defensive and offensive actions.

26. In a defensive scenario, given the lack of time to respond to missile attacks, for example, and in the interest of protecting combatants and especially civilians, some of the critical functions must be done autonomously. In these situations, greater flexibility may be granted to AWS.

27. On the other hand, at offensive scenarios, greater levels of human control and limited autonomy on critical functions should be mandatory in combat situations with deployment of AWS, especially in populated areas.

Working definitions for a legal framework

28. This paper is linked to the other Brazilian contribution to the GGE presented under the title *Operationalizing the Guiding Principles: a roadmap for the GGE on LAWS* [2].

29. In that document, Brazil proposes paths of action leading to advancements in the governance of LAWS, ultimately arriving at the codification of specific International Humanitarian Law rules and a new protocol under the CCW.

30. Such a protocol could establish the general obligation of maintaining meaningful human control over the use of force through the activation of AWS, as well as specific obligations regarding control over critical functions of selecting and engaging targets. Furthermore, specific categories of AI weapons should be prohibited on the basis of ethical and moral considerations.

31. The working definitions presented in this paper are designed to contribute to the drafting of that legal framework.

[2] Available at: https://documents.unoda.org/wp-content/uploads/2020/08/CCW-GGE.1-2020-WP.3-.pdf