Group of Governmental Experts 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

24 August 2018

English only

Second Session
Geneva, 27 - 31 August 2018
Item 6 of the provisional agenda
Other matters

Categorizing lethal autonomous weapons systems - A technical and legal perspective to understanding LAWS

Submitted by Estonia and Finland

I. Introduction

1. The aim of this paper is to **facilitate the discussion on the conceptualization and categorization of lethal autonomous weapons systems** (LAWS) further while focusing on the critical functions in the targeting cycle. To make progress in the discussion on LAWS, there is a clear need to develop shared and commonly understood basic concepts. However, due to the many effects and implications of advanced machine autonomy in association with weapon systems, even a working definition of LAWS is very difficult to frame. To complicate matters further, each word in the scoping phrase "lethal autonomous weapons systems" may require some reflection and clarification:

- **Lethality** – There is no clear reason to exclude less-than-lethal weapons from the discussion – lethality is not a defining feature of any weapon system, autonomous or otherwise. An instrument that is intended to cause less-than-lethal injuries to persons, or harm to objects, is nonetheless a weapon. Also, a weapon intended to be less-than-lethal may well prove to be lethal in certain circumstances.
- **Autonomy** – Not only is there a need to understand the meaning of autonomy, it may be necessary to elaborate on different dimensions and degrees of autonomy. While a reference to *full* autonomy may at first seem convenient for categorizing weapon systems, there is no technological reference point when a system becomes *fully* autonomous.
- **Weaponry** – One can easily get stuck on the idea of projectiles (such as bullets or missiles) in the context of weapons. However, other capabilities, such as lasers, high power microwave (HPM), nanoparticles, or other mechanisms, could potentially be used to cause harm to an adversary.
- **Systems** – Autonomous systems cover a wide spectrum. Despite their structure or appearance, only those aspects that have a direct and concrete link to the process of projecting force (flow of information, decision making and timing) are relevant in understanding the challenges for human control.

II. Characteristics of machine autonomy

2. To create results that can withstand time, the discussions on LAWS must reflect the undeniable direction of technological development. The development of artificial intelligence (AI) should be seen as a logical progress in computing science. The level of abstraction of computing keeps getting higher and higher, leading towards increasing possibilities for various levels of machine autonomy. Past experience has shown that once new technology proves to work, society quickly adopts it, and later its use becomes the accepted norm.

3. **As a result of technological developments, the way humans use machines and interact with them is changing.** In complex systems the human role will have various postures in relation to the machine. The underlying fundamental question is about human conduct, and its limits, in relation to machines. As a part of this development, warfare is only one, although in many ways the ultimate, application of autonomous technology. The developments in the civilian sector may easily surpass military capabilities.

4. As already mentioned, the notion of "fully autonomous" is problematic, as autonomy is always a relative term. To facilitate the discussion and understanding of the nature of AI-based machine autonomy, it could be helpful to make a **difference between (1) automation, (2) autonomy, and (3) independence.**

1. Automation, automated functioning

5. Basically, automation as a concept means known, predictable pre-programmed responses in any situation in a defined task. To be reliably able to cope with any imaginable situation occurring in a complex real-time environment, an infinite number of various states of a system should be foreseen and defined. Therefore, fully deterministic automation is only feasible in rather simple and limited cases. Pattern matching algorithms which are used for target selection in many weapons systems are very dependable and robust in practice but contain the element of probabilistic behavior.

6. When the overall complexity of any system reaches a certain point, the systems level behavior will be stochastic, involving a degree of randomness or chance, even if the individual subsystems and components alone are fully deterministic. **The assumption that automation (in contrast to autonomy) would always produce inherently or structurally stable and thereby safe systems behavior, is not valid in complex systems.** Rather, the controlled and stable behavior of complex automated systems is an achievement of thorough systems design and rigorous testing.

2. Autonomy, autonomous functioning

7. As a more advanced solution, **autonomy should be understood as a capability to perform the given task(s) in a self-sufficient and self-governing manner**. This includes the freedom of self-planning in the tasks and required subtasks. The programming and control structures behind AI systems are fundamentally about task execution. Complex behaviors are abstracted and divided into simple tasks on various levels. The tasks are run and activated in parallel and in sequences. Again, testing and experience of use are in a crucial role to achieving reliable and predictable systems behavior.

8. Depending on the situation, tasks vary in duration and complexity. Each task has its limitations, pre-conditions, terms, rules of engagements, etc. In order to carry out the tasks, advanced AI and cognitive features are required in the implementation of complex behaviors. Paradoxically, an autonomously functioning system may, while executing a task, communicate interactively with humans or other entities of the system. For instance, a system capable of monitoring and self-assessing the limits of its authorization, or asking for advice, is cognitively much more sophisticated than one that is not able to conduct context sensitive interaction or behavior.

9. Importantly, autonomy is not an on/off feature, so instead of "autonomous systems" it would be better to use the expression "systems having autonomous features or functions". It is very difficult to define the level or degree of autonomy; as various systems are very different by nature. In the LAWS context, it is important to keep the focus on the targeting cycle and the conditions of the authorization to use lethal force. Of particular importance in the tasking are the delay from command to execution, i.e., understanding the dynamics of the task, and the time window of authorization.

3. Independence, independent functioning

10. In contrast to autonomous operation, only true independence, however, means that the system would be capable of defining and thereby deciding the ultimate goals of its functioning, very much like humans do. The independent targeting capability would follow as a subordinate behavior to the system's own self-motivation. Of course, this is undesirable but also highly unlikely in the foreseeable future as it would require human-like or superhuman AI, available beyond the singularity point.

III. Human-weapon interaction

11. **Humans must retain ultimate control over decisions of life and death**. While this proposition may be viewed as a moral imperative, it also has a legal basis. In an armed conflict, rules and principles of international humanitarian law place restrictions on the use of violence by limiting the choice of means and methods of warfare, and by protecting those not (or no longer) taking part in hostilities. These restrictions apply to the conduct of individual humans as well as to the conduct of States. As abstract entities, States can only comply with international law through the acts of their agents – that is to say, humans whose actions or omission are attributable to States.

12. Thus, either directly or indirectly, the focus of international humanitarian law is the conduct of humans in armed conflict. Weapons, including weapon systems with some degree of autonomy, are instruments that humans choose to use in the conduct of hostilities. It is for the humans to ensure that the instruments that they choose are capable of being used consistently with the law, and are in fact so used. This plainly presupposes that humans can exert a degree of influence over the operation of the weapons. In other words, the duty to reason, rationalize, make decisions, and therefore carry the ultimate responsibility of the outcome of the use of a weapon remains always with the human operator. An increase in the capacity of AI or advancement in the level of autonomy in a weapons system does not change this fundamental principle.

13. **The need to exercise human control over weapons does not arise from a discrete rule of international law,** whether existing or emerging. Rather, human control over weapons is an important and likely indispensable way for humans to ensure that the use of violence complies with international law. Therefore, as a minimum requirement under international humanitarian law, **humans must exercise such control over a weapon system as may be necessary to ensure the operation of that weapon system consistently with international law.**

14. **Human control over weapons can be exercised in various ways and at various times**. Most obviously, humans can exert influence on the operation of a weapon system by operating it 'manually' – for example, by aiming a rifle at a target and pulling the trigger. In such circumstances, the human has very extensive control over the weapon and consequently also bears the bulk of the responsibility for any failure to comply with the applicable law. However, other means of human control will also contribute to the lawful use of the weapon. For example, human-controlled design and manufacturing processes impact on whether the bullets fired from the rifle comply with the prohibition of superfluously injurious weapons.

15. With respect to technologically more sophisticated weapon systems, the design and manufacturing processes (as well as testing) play a comparatively greater role in ensuring compliance with the law. For example, with precision-guided projectiles, the design

features are at least as important as the conduct of the operator when it comes to ensuring respect for the principle of distinction.

16. In light of this dynamic, it becomes **difficult to provide a technical statement of meaningful human control.** The kind and degree of human control that must be exercised at various points leading up to and including the use of the weapon depend heavily on the nature of the weapon and circumstances of its use. Clearly, however, if the ability of the operator to exert control over the weapon is restricted – e.g. because of the autonomous capabilities of the weapon system – the designers and manufacturers must exert more control and bear greater responsibility. In any event, it is the combination of all human interventions that must ensure that any application of force by the weapon complies with international humanitarian law, rather than each of them individually.

17. That said the most critical point of human control is the final interaction between a human a weapon system before that system delivers force. With respect to a system with a significant degree of autonomy, this might be the point at which the system is activated or switched to autonomous operation. At that point, the operator of the system must reassure him/ herself that the system would, under the circumstances, operate consistently with the law. This would require, as a minimum, an understanding of the performance characteristics of the system and of the operational environment. If the operator lacks such an understanding, or based on that understanding has no confidence as to compliance with the law, he/she must not permit the weapon to deliver force.

18. Thus, before an operator deploys a weapon system with a significant degree of autonomy, he/she must make sure that the area of operations remains sufficiently clear of persons and object that cannot be targeted or that the system is capable of sufficiently distinguishing between different persons and objects. Beyond this general point, the examination of the matter turns unavoidably case-specific. However, as a reliability, transparency and accountability measure (i.e. to detect errors, investigate malfunctions, and address the potential attribution problems), it may be advisable for a system with autonomous functionality to comprehensively log its activity.

19. To be meaningful, **human control does not necessarily have to be exercised contemporaneously with the delivery of force**. For example, it is possible to have meaningful human control over non-command-operated anti-vehicle land mines. With respect to such weapons, human control could be exercised through various design features (e.g. self-destruction or self-neutralization mechanism), decisions relating to the emplacement and other precautionary measures (e.g. warning the civilian population). An appropriate combination of such measures can ensure the compliance of the use of such a weapon, and potentially a more sophisticated autonomous weapon, with the law.

20. Finally, **meaningful human control cannot always require the technical capability to cancel an attack that has already commenced**. Unguided projectiles (such as rifle bullets or artillery shells) and simple guided projectiles (such as heat seeking missiles) cannot be stopped once launched. Yet the use of such projectiles would generally not be seen as lacking meaningful human control.

21. From the point of view of compliance with IHL, in particular the principles of distinction, proportionality and precaution, the operator and the commander should be guided by the following:

- proceed only after a comprehensive contextual risk assessment,
- commit a limited authorization to use lethal force with regard to the assigned military task, in a defined time window and conditions, and
- assign the task only to a particular system known to be capable of performing the mission with the required briefing/parametrization based on relevant a priori information and policies.

IV. Conclusion

22. In summary, this working paper makes the following observations:

- Humans must always retain ultimate control over decisions of life and death, and to ensure that the operation of weapons is consistent with international law.
- Computing is developing towards increasing possibilities of various levels of machine autonomy. Past experience has shown that once new technology proves to work, society quickly adopts it, and later its use becomes the accepted norm.
- The distinction between automated and autonomous functioning is not clear-cut. This is partly because both automated and autonomous systems can have a degree of unpredictability, therefore controlled and stable behavior of any complex system must be achieved by means of thorough systems design and rigorous testing.
- Because of the ongoing development and seamless shift from automation to more autonomous functioning, it would be problematic to set different standards or requirements for existing weapon systems and for new types of systems.
- Now and in the foreseeable future, both automated and autonomous systems perform tasks assigned by human operators, who bear ultimate responsibility for the use of the systems.
- The requirements of a military command chain emphasize the nature of task execution; understanding of the time dimension (delay) and the dynamics of the situation are crucial in the task definition and authorization of lethal force.

Annex

Analogy from civilian regulations – Case: EU Machinery Directive

1. As an example of an analogy from civilian industry, where efforts have been made to regulate the safety measures of various machines, the *European Union Machinery Directive* (2006/42/EC, edition 2.1, July 2017) (EUMD) is briefly reviewed. The directive is intended as a unified code of conduct to promote safety and free commerce in its domain by harmonizing common rules for national administrative organizations and safety surveillance authorities. The applicable scope of the EUMD vary from hand-held tools to heavy work machines, such as agricultural, forestry, or construction machines, etc. The focus and objective of EUMD has been to address safety and responsibility. Despite the fact that it has particularly been expressed not to be applicable to weapons, it serves as an interesting reference relevant to LAWS.

2. The leading principle of the EUMD is to trust the manufacturer's responsibility and assurance of the products' conformity with the given standards. However, certain listed categories of potentially dangerous machines, or machine types without existing standards, require a type examination by the notified inspection authority in any case. In manual machinery, most of the responsibility stays on the user/driver, who has the contextual responsibility of the use-case. When the technical complexity and automated functions increase, the more responsibility falls to the manufacturer.

3. EUMD does not yet deal with machine autonomy, other than defining emergency stop mechanisms, and the detection of humans. Responsibility on the human safety in proximity of work machines can be handled in various ways, or as combined: by denying human access in certain work areas, relying on onboard sensors to detect and track humans or any movement in certain perimeters or a rule-based solution such as traffic rules. At the moment there exist standards of safety sensors for human detection (in a dangerous zone or proximity to a work machine) for indoor use, but not yet for outdoor use. Autonomously driving road cars operate under traffic laws in a different regulative framework to detect pedestrians. So far in most countries they operate within specific testing permissions or under driver's monitoring and responsibility.

4. Reliable and qualitative/distinctive human detection is a key enabler for all kinds of everyday robotic applications and autonomous devices. Both primary sensors (laser scanners, millimeter-wave radars, hyperspectral imaging, etc.) and signal processing algorithms for human detection in various environments, conditions, and within different legislative regimes have been under massive research and development efforts. Therefore, one can expect to see major progress in this domain sooner than later. The expected impact on the debate on LAWS will likely be immediate.