Commentary on the Guiding Principles

International Panel on the Regulation of Autonomous Weapons, September 2020

~ original text of the Guiding Principles in bold letters ~

The International Panel on the Regulation of Autonomous Weapons welcomes the opportunity to comment on the Guiding Principles[1]. We derive our remarks from previous iPRAW publications with the intent to contribute to an operationalization of the principles and to offer comments to the normative and operational framework.

It was affirmed that international law, in particular the United Nations Charter and International Humanitarian Law (IHL) as well as relevant ethical perspectives, should guide the continued work of the Group. Noting the potential challenges posed by emerging technologies in the area of lethal autonomous weapons systems to IHL, the following were affirmed, without prejudice to the result of future discussions:

REMARKS ON THE DEFINITION OF LAWS: iPRAW recommends regarding the term lethal autonomous weapon systems as shorthand for various weapon platforms as well as systems of systems with machine 'autonomy' in the functions required to complete the targeting cycle. This stands in stark contrast to a categorical definition of LAWS. A categorical definition drawing on technical characteristics in an effort to separate "LAWS" from "non-LAWS" is unable to account for the already existing plethora of systems with autonomous functions and could, as technology progresses further, never be future-proof because almost every conceivable future weapon system can optionally be endowed with various autonomous functions. Even more importantly, in the CCW context a technical definition of a category of weapons would miss the point: while technologies like data-driven computational methods (i.e. artificial intelligence, machine learning) do enable many autonomous functions, the legal, ethical, and operational challenges ascribed to LAWS arise not from particular technologies but from a potential lack of human involvement. Hence, if CCW states parties want to define LAWS, a technology-agnostic conceptualization as presented above, with a focus on 'autonomous' machine functions (instead of specific units or platforms) and human-machine-interaction is the most suitable and prudent approach. The ICRC presented a broad notion in this vein already: "*Any weapon system with autonomy in its critical functions—that is, a weapon system that can select (search for, detect, identify, track or select) and attack (use force against, neutralize, damage or destroy) targets without human intervention.*"[2] iPRAW recommends following this train of thought and fleshing it out further, for instance by excluding existing systems like stationary, anti-materiel weapons used solely to defend against incoming munitions.

1 Guiding Principles affirmed by the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, CCW/MSP/2019/9, p. 10.

2 ICRC (2016), *Views of the International Committee of the Red Cross on Autonomous Weapon Systems*, p. 1.

(a) International humanitarian law continues to apply fully to all weapons systems, including the potential development and use of lethal autonomous weapons systems;

IHL: iPRAW agrees that international humanitarian law applies to the development and use of all weapon systems and calls for human control before and during attack. Precaution during attack remains feasible when the operator/commander has sufficient situational understanding and options for intervention along the lines of iPRAW's concept of human control (details are discussed below under Principle c).

RELEVANCE OF OTHER FIELDS OF LAW: Another strong incentive for the application of human control in the use force is the ethical and legal principle of **human dignity**. It is inter alia enshrined the principles of humanity of the Martens Clause, which is part of the CCW preamble, as well as in international human rights law. Those legal sources contribute to the development process of the normative framework mentioned in the GGE's mandate.

(b) Human responsibility for decisions on the use of weapons systems must be retained since accountability cannot be transferred to machines. This should be considered across the entire life cycle of the weapons system;

ROLE OF THE LIFE CYCLE: The design and all other steps of the weapon system's life cycle are enablers to allow for human control during attack.

ATTACK: To unpack the concept of human control, a common understanding of what constitutes the start of an attack can be useful. In this context, the most relevant point in the mission thread is not defined by the launch or activation, but by the final necessary decision necessary for target selection and engagement by a human. Weapon systems with autonomous functions potentially move the final human decision to a very early stage of the operation. With regard to the legal obligation to abide by the various principles of IHL this effect could be challenging for two reasons: First, it can increase the level of abstraction in the target selection process (i.e. class of targets instead of specific target). Second, the environment might change during this extended timespan between targeting decision and engagement, e.g. outdating the initial proportionality assessments. The underlying notion of attack will therefore influence the understanding of the concept of human control in a regulation of autonomous weapon systems. This is because IHL principles like distinction and proportionality are legally required during the planning phase, but, to a certain extent, become a question of feasibility in attack. This would alter the need or necessary level of human control in attack.[3]

[3] For details see iPRAW (August 2019), *Focus on Human Control*, p. 5-6.

(c) Human-machine interaction, which may take various forms and be implemented at various stages of the life cycle of a weapon, should ensure that the potential use of weapons systems based on emerging technologies in the area of lethal autonomous weapons systems is in compliance with applicable international law, in particular IHL. In determining the quality and extent of human-machine interaction, a range of factors should be considered including the operational context, and the characteristics and capabilities of the weapons system as a whole;

HUMAN CONTROL: The 'human element' can be conceptualized as human control, meaning e.g. the requirement for situational understanding by the human and the option to intervene built-in by design and available any time during use. The concept covers the whole life cycle of a weapon system.[4]

In iPRAW's understanding, **human control does not necessarily equal direct manipulation**. The directness of the means whereby the agent seeks to control some object is related only contingently to the degree of control. Under some circumstances, manipulation that is more direct enables greater control, while in other circumstances the presence of an intervening mechanism might be the better option to reach the desired outcome. The increasing number of assisting systems does not necessarily increase precision, though, as they make the weapon system also more complex and possibly less predictable. A prudent balance of operational needs and situational understanding is crucial.

Control is also not to be understood as a singular event during or at the end of the targeting process, but as a process, that requires at least a frequent understanding of the situation. The adequate type and level of human control depends on the individual operational context.

OPERATIONAL CONTEXT: To account for a multitude of battlefield applications a regulation of LAWS might have to remain rather abstract with regard to the type and level of human control. It could encompass, however, all steps in the weapon's life cycle. Those could be, for example, design requirements, training, rules of engagement as well as the explicit call for human control in the targeting process and during the actual attack.

The operational context is crucial for defining the necessary type and level[5] of human control and multiple factors contribute to the determination of what level of human control is adequate in a given situation. A **'one-size-of-control-fits-all' solution** that addresses all concerns raised by the use of autonomous weapon systems will thus most likely not be achievable. In the context of the CCW, the most relevant aspect would be the **risk for violations of IHL (due to a lack of situational understanding or timely intervention)**. One option to address this would be a classification of factors that define the operational context in order to derive consequences for the implementation of human control.[6] The shortcoming of such a kind of typology lies in the multitude of combinations of environmental factors, operational requirements, and weapon capabilities it cannot account for. Instead, a regulation would be more useful if it included general approximations to be specified in each case along the lines of existing IHL considerations. Best practices and other dynamic soft law measures could accompany a more abstract regulation. In addition, tabletop exercises[7] and other scenario-based workshops[8] could facilitate a better understanding of the context before actually fielding weapon systems with autonomous functions.

---

4 For details see iPRAW (March 2018), *Focus on the Human-Machine Relation in LAWS*.

5 Not: the need for human control.

6 For an exploratory approach to define relevant criteria see: Marcel Dickow et al. (2015), *First Steps towards a Multidimensional Autonomy Risk Assessment (MARA) in Weapons Systems*.

7 E.g. by UNIDIR: Giacomo Persi Paoli (April 2020), *Human Element in the Decisions about the Use of Force*, p. 8.

8 iPRAW is currently developing a series of events starting in October 2020.

(d) Accountability for developing, deploying and using any emerging weapons system in the framework of the CCW must be ensured in accordance with applicable international law, including through the operation of such systems within a responsible chain of human command and control;

ACCOUNTABILITY, PREDICTABILITY AND CONTROL: iPRAW recommends that the leading characteristic of the human-machine interaction should be that of human control and machine dependence on humans in the execution of the targeting cycle. The control exercised by the operator must be sufficient to reflect the operator's intention for establishing legal accountability and ethical responsibility for all ensuing acts.

Legal as well as operational considerations show the necessity of human control over weapon systems with the requirement for predictability through a strong human involvement as one common denominator. In addition to that, a philosophical perspective on control helps to define this abstract concept, showing that reliability and predictability determine the level of control that humans must ascertain over objects.

To allow for predictability and to abide by legal requirements, the human operator must be aware of the state of the system as well as its environment. Therefore, the system's design must allow the operator to monitor both. This can be achieved through frequent (technical or operational) points of inquiry throughout the targeting cycle. In addition to this situational understanding, the human operator needs options to interact with the system.

(e) In accordance with States' obligations under international law, in the study, development, acquisition, or adoption of a new weapon, means or method of warfare, determination must be made whether its employment would, in some or all circumstances, be prohibited by international law;

ARTICLE 36 imposes an obligation on states to ensure the use of weapons, means or methods of warfare is lawful prior to making new systems operational. It is left to states to determine their own domestic processes for evaluation. **Article 36 is necessary but not sufficient:** First, not all states have robust procedures and mechanisms on the legal review of new weapons. Second, it requires only a determination that weapons do not violate IHL (and possibly international human rights law) in general, a fairly low threshold to meet since just one IHL conform application is sufficient. Given the increasing innovation of weapon systems, it may become more and more difficult for a commander to understand how a system works and evaluate whether it will be lawful to use it in a given situation absent a supplemental review or process.

Moreover, the testing and evaluation of systems with data-driven computational methods presents several challenges, which may translate to reviews that include incomplete information or cannot quantify the reliability of the system. While Article 36 legal reviews of new weapons remain important, there is urgent need for additional processes or guidance such as those recommended by Boulanin/Verbruggen,[9] aimed at making the legal review process more robust and fit for purpose. The notion of human control could be added as legal principle for consideration when conducting legal reviews.[10] The challenge however is to universalize the practice of weapon reviews and make it more transparent.

---

9 See Vincent Boulanin & Maaike Verbruggen (2017), *Article 36 Reviews: Dealing with the Challenges Posed by Emerging Technologies*, SIPRI, pp. 22-25.

10 See e.g. Thompson Chengeta (2019), Is existing law adequate to govern autonomous weapon systems?, <http://ceur-ws.org/Vol-2540/FAIR2019_paper_9.pdf>- p. 3-4.

(f) When developing or acquiring new weapons systems based on emerging technologies in the area of lethal autonomous weapons systems, physical security, appropriate non-physical safeguards (including cyber-security against hacking or data spoofing), the risk of acquisition by terrorist groups and the risk of proliferation should be considered;

-

(g) Risk assessments and mitigation measures should be part of the design, development, testing and deployment cycle of emerging technologies in any weapons systems;

MITIGATION MEASURES: Given the high stakes of failure in such a system, it is imperative to design the weapon system to fail gracefully and along defined procedures. For some critical functions, it even needs to be designed in such a manner that the specific function can be fulfilled only by handing over the control to a human being. To allow for a handing-over procedure, a robotic system needs to have defined and comprehensible modes of autonomous functionalities. Otherwise, humans will not be able to regain control in (time) critical situations as they do not understand the system's operational mode and will lack contextual awareness. Control is thus a necessity for the implementation of the human-machine-handshake protocol. Since even well-designed failure modes bear a certain risk, the choice of autonomous functions needs to be considered carefully in the first place.

Migrating from a mode A to a higher mode of autonomy B should be subject to direct human intervention: Cases where a person fails to perform critical functions of a mission and the machine takes over (similar to the anti-lock system in cars) have to be designed very carefully to avoid undermining this principle.

The design of failure modes in all stages of the targeting cycle must allow for enough time and information for situational understanding. That would include a clear indication of responsibilities (What is demanded from the operator? What can the machine still do on its own?) and an immediate halt on the use of force.

(h) Consideration should be given to the use of emerging technologies in the area of lethal autonomous weapons systems in upholding compliance with IHL and other applicable international legal obligations;

-

(i) In crafting potential policy measures, emerging technologies in the area of lethal autonomous weapons systems should not be anthropomorphized;

TERMINOLOGY: iPRAW uses the term **computational methods** to refer to the technology and techniques used to enable machine autonomy. The panel recommends using this term in place of artificial intelligence to avoid the bias that occurs when technology is described in anthropomorphized terms, which run the risk of evoking unrealistic notions of capabilities or unnecessary philosophical debates.

When mentioning algorithms that sense, recognize, plan, decide, or act autonomously, we do not mean to anthropomorphize machines. Instead, these terms should be understood as shorthand descriptions.

(j) Discussions and any potential policy measures taken within the context of the CCW should not hamper progress in or access to peaceful uses of intelligent autonomous technologies;

-

(k) The CCW offers an appropriate framework for dealing with the issue of emerging technologies in the area of lethal autonomous weapons systems within the context of the objectives and purposes of the Convention, which seeks to strike a balance between military necessity and humanitarian considerations.

-