**Convention on Certain Conventional Weapons (CCW)**

**Group of Governmental Experts on Lethal Autonomous Weapons Systems**

**25-29 March 2019, Geneva**

# Statement of the International Committee of the Red Cross (ICRC)

## under agenda item 5(b) further consideration of the human element in the use of lethal force; aspects of human-machine interaction in the development, deployment and use of emerging technologies in the area of lethal autonomous weapon systems

In our last intervention, the ICRC explained the legal basis for human control, that is, the limits on autonomy that can be deduced from IHL rules.

In the ICRC's view, ethical considerations also demand human control over weapon systems and the use of force.

As you know, the ICRC is convinced that human intention and agency – the human actor – must be sufficiently retained in decisions to use force.

Moral responsibility for these decisions cannot be delegated to a machine, no more than legal responsibility can.

With the legal and ethical bases of human control in mind, the ICRC wishes to dedicate its intervention this afternoon to the key elements of human control demanded by legal and ethical considerations (noting that ethical concerns might demand additional limits on autonomy – an issue we have spoken about before).

These key elements are:

1. Human supervision and the ability to intervene and deactivate
2. Predictability and reliability, and
3. Operational constraints.

### 1. Human supervision, and ability to intervene and deactivate

**Human supervision, and the ability to adapt to changing circumstances, is essential to ensure compliance** with IHL, including when carrying out an attack with a weapon system with autonomy in its critical functions. This requires supervision of both the weapon system and the target area, in other words **situational awareness**.

Generally speaking**, constant human supervision of the weapon system and the target area may be required**, so that the operator has sufficient information and understanding about the operation of the weapon system, the environment of use, and the interaction of the two, over the given time period and geographical area. This information is necessary for making the context-based legal judgments required by IHL.

A **physical and/or communication link** that permits adjustment of the engagement criteria and the ability to cancel the attack, as well as sufficient time for such intervention, **is necessary if the supervision is to serve its purpose of ensuring compliance with IHL**.

Without human supervision – and ability to intervene and deactivate – it is difficult to envisage how operators/commanders could take into account changes in the situation and thereby exercise the legal judgements required by IHL in carrying out attacks.

IHL requires the attacker to maintain awareness of, and adapt to, continuously changing circumstances, even after the decision to attack. A party to the conflict must do everything feasible to cancel or suspend an attack if it becomes apparent that the target is not a military objective, or that the attack may be expected to cause incidental loss of civilian life, injury to civilians, damage to civilian objects, or a combination thereof which would be excessive in relation to the concrete and direct military advantage anticipated (API article 57(2)(b), CIHL rule 19).

The practical effect of this rule is that weapon systems must remain under human supervision, and must permit the user to, where feasible, cancel, suspend or modify attacks, up until the execution of the attack (or halting of the attack).

Feasibility means what is practicable or practically possible, taking into account all the circumstances at the time, including humanitarian and military considerations.

The existence of autonomy in the critical functions of a weapon system may render precautionary measures unfeasible.

Precautions will not always be precluded when using an autonomous weapon system, but the ICRC is interested to hear from States about how they see the interaction between autonomy and feasibility and precautions.

In the ICRC's view, the use of an autonomous weapon system that does not permit the taking of precautions such as cancelling or suspending an attack – in situations where there is a reasonable likelihood that the circumstances will change enough to render an attack unlawful – would likely be unlawful.

## 2. Predictability and reliability

**Autonomous weapon systems, since they self-initiate attacks, all raise concerns about unpredictability**, owing to varying degrees of uncertainty about **location, timing and/or nature of the subsequent machine-initiated attack**.

**Predictability** (knowledge of the consequences of use) **and reliability** (likelihood of failure) are dependent on the: [1] weapon system design (including the software and algorithms that control the system); task it is used for; nature of the environment where it is used; and interaction between the system and the environment**.**

**All autonomous weapon systems**, which operate based on interaction with their environment, **raise questions about human control and predictability**. The greater the complexity of the environment and complexity of the task, the greater the need for human control and the less tolerance of autonomy, from both a legal and ethical perspective.

[1] **Predictability** is the ability to "say or estimate that (a specified thing) will happen in the future or will be a consequence of something". Applied to an autonomous weapon system, predictability is knowledge of how it will function in the circumstances of use, and the effects that will result. **Reliability** is "the quality of being trustworthy or performing consistently well". In this context, reliability is knowledge of how consistently the machine will function as intended, e.g. without failures or unintended effects.

**Humans can exert some control over autonomous systems through "human on the loop" interaction** with them. However, this is not a panacea due to human-machine interaction problems, such as automation bias, over-trust in the system, or lack of operator awareness of the system state at the time of intervention.

Further, there are challenges in **quantifying the level of predictability** (and reliability) **needed to ensure a level of human control, or judgement, sufficient for IHL compliance (and ethical acceptability)**. Testing also raises unique challenges since it is not possible to test all possible environmental inputs to an autonomous system.

In order to comply with IHL, those who plan or decide on an attack using an autonomous weapon system must understand its capabilities and limitations in the given circumstances, in order to determine whether it will perform lawfully in the given circumstances. This also requires knowledge of the environment over time (see above on human supervision).

In general, IHL demands a high level of certainty about the prevailing circumstances and the effect of the chosen means and methods of warfare. One example of this is the rule on the loss of civilian protection in attack. During an attack, doubt as to status of a person must be resolved in favour of treating the individual as a civilian.

It is challenging to pinpoint exactly where IHL lines will be crossed with respect to predictability and reliability. The ICRC would be interested to hear from States what levels of predictability/reliability are demanded by IHL. Indeed, the absence of clarity on this issue is one of the reasons that we have called for States to set internationally agreed limits on autonomy in weapon systems.

**A weapon system that is unpredictable by design would be unlawful by its nature**.

Let us explain what we mean by that.

**Unpredictability of a weapon system's interaction with its environment will inevitably create risks for protected persons and objects** in that environment.

There are fundamental **concerns about autonomy in critical functions controlled by machine-learning algorithms, since these are generally unpredictable** in their functioning, **not transparent** (i.e. they are "black boxes" and their functioning cannot be explained), and their performance therefore cannot be verified during testing. Not only are such algorithms unpredictable, but they **can also introduce bias**, whether by design or through bias in the data used to "train" (develop) the algorithms.

This kind of 'unpredictability by design' would raise concerns in all situations, as the person who plans or decides on an attack cannot have reasonable certainty about the effects of the weapon.

Any **weapon systems with autonomy in critical functions whose software can set its own goals, or learn, change or adapt its functioning after deployment, would be inherently unpredictable, and therefore beyond human control and unlawful under IHL**. This is because any assessment that the user has made at the moment of activating the weapon to ensure compliance with IHL would immediately become invalid after deployment. **Self-organising swarms used as autonomous weapons could raise similar concerns** due to their inherent unpredictability of their behaviour in the environment.

## 3. Operational constraints

**Operational parameters and constraints are important for human control** in particular the: **task** the weapon is used for; **types of targets** it attacks, **type of force** (and effects) it employs; **operating environment**; **duration of autonomous operation** (time-frame); and **scope of movement over and area** (mobility).

An autonomous weapon system must be capable of being used – and must be used – in accordance with existing rules of IHL - notably the rules of distinction, proportionality and precautions in attack, which require complex context-based assessments by commanders based on the circumstances prevailing at the time of the attack. As the ICRC explained in our intervention this morning, these assessments must be reasonably proximate in time to the attack (or "strike") to comply with these rules.

**Autonomy in the critical functions of weapons complicates the ability of the commander, to make these context specific judgements** since, upon activation, they do not know the timing, location and nature of the subsequent attack(s) that the weapon will self-initiate. The lawfulness of use relies on the continuing validity of IHL assessments and planning assumptions made at the point of activation.

Whether an autonomous weapon system will operate within the constraints of IHL once activated will depend on the technical performance of the specific weapon, especially its predictability and reliability (see above), and the specific circumstances and environment of use. Predictability in the consequences of its use will depend not only on the technical design of the system, but on variations in the environment over time and the interaction of the system with that environment, taking into account the task it is used for.

The more variable the environment of use, or the longer the timespan between the human decision to activate the weapon and the eventual use of force initiated by the weapon system in response to the environment, the greater the risk of IHL violations.

The risk that IHL might be violated can be reduced by manipulating operational parameters like the environment in which the weapon system is to operate, the mobility of the weapon system in space and the time-frame of its operation, in order to increase predictability in the consequences of its use. The more predictable (non-dynamic) the environment, and the more highly constrained the system is in time and space, the greater predictability in the consequences of activating the system.

The nature of the spatial/temporal limitations required by IHL is one on which further clarity would be useful. Reaching agreement on the kinds of operational constraints necessary as part of human control for IHL compliance (and ethical acceptability) could be a part of States' efforts to set limits on autonomy in weapon systems.

However, given the high degree of unpredictability of most real world conflict environments, it is likely that operational constraints alone will only help avoid an unacceptable risk of IHL violations in the narrowest of circumstances, and will generally not be sufficient to ensure IHL compliance with IHL in carrying out an attack with an autonomous weapon system.

Mr Chair, this intervention has already been rather long, so we will conclude there and look forward to contributing to the discussions in the coming days.

Thank you.