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

20 August 2019

English only

Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems
Geneva, 25–29 March 2019 and 20–21 August 2019
Agenda item 5
Focus of work of the Group of Governmental Experts in 2019

Autonomy, artificial intelligence and robotics: Technical aspects of human control

Submitted by International Committee of the Red Cross (ICRC)[1]

1. The International Committee of the Red Cross (ICRC) has emphasized the need to maintain human control over weapon systems and the use of force, to ensure compliance with international law and to satisfy ethical concerns. This approach has informed the ICRC's analysis of the legal, ethical, technical and operational questions raised by autonomous weapon systems.

2. In June 2018, the ICRC convened a round-table meeting with independent experts in autonomy, artificial intelligence (AI) and robotics to gain a better understanding of the technical aspects of human control, drawing on experience with civilian autonomous systems. This report combines a summary of the discussions at that meeting with additional research, and highlights the ICRC's main conclusions, which do not necessarily reflect the views of the participants. Experience in the civilian sector yields insights that can inform efforts to ensure meaningful, effective and appropriate human control over weapon systems and the use of force.

3. Autonomous (robotic) systems operate without human intervention, based on interaction with their environment. These systems raise such questions as "How can one ensure effective human control of their functioning?" and "How can one foresee the consequences of using them?" The greater the complexity of the environment and the task, the greater the need for direct human control and the less one can tolerate autonomy, especially for tasks and in environments that involve risk of death and injury to people or damage to property – in other words safety-critical tasks.

4. Humans can exert some control over autonomous systems – or specific functions – through supervisory control, meaning "human-on-the-loop" supervision and ability to intervene and deactivate. This requires the operator to have:

- situational awareness
- enough time to intervene
- a mechanism through which to intervene (a communication link or physical controls) in order to take back control, or to deactivate the system should circumstances require.

[1] To download the full report visit: https://www.icrc.org/en/war-and-law/weapons/ihl-and-new-technologies

5. However, human-on-the-loop control is not a panacea, because of such human-machine interaction problems as automation bias, lack of operator situational awareness and the moral buffer.

6. Predictability and reliability are at the heart of discussions about autonomy in weapon systems, since they are essential to achieving compliance with international humanitarian law and avoiding adverse consequences for civilians. They are also essential for military command and control.

7. It is important to distinguish between: reliability – a measure of how often a system fails; and predictability – a measure of how the system will perform in a particular circumstance. Reliability is a concern in all types of complex system, whereas predictability is a particular problem with autonomous systems. There is a further distinction between predictability in a narrow sense of knowing the *process* by which the system functions and carries out a task, and predictability in a broad sense of knowing the *outcome* that will result.

8. It is difficult to ensure and verify the predictability and reliability of an autonomous (robotic) system. Both factors depend not only on technical design but also on the nature of the environment, the interaction of the system with that environment and the complexity of the task. However, setting boundaries or imposing constraints on the operation of an autonomous system – in particular on the task, the environment, the timeframe of operation and the scope of operation over an area – can render the consequences of using such a system more predictable.

9. In a broad sense, all autonomous systems are unpredictable to a degree because they are triggered by their environment. However, developments in the complexity of software control systems – especially those based on AI and machine learning – add unpredictability in the narrow sense that the process by which the system functions is unpredictable.

10. The “black box” manner in which many machine learning systems function makes it difficult – and in many cases impossible – for the user to know how the system reaches its output. Not only are such algorithms unpredictable but they are also subject to bias, whether by design or in use. Furthermore, they do not provide explanations for their outputs, which seriously complicates establishing trust in their use and exacerbates the already significant challenges of testing and verifying the performance of autonomous systems. And the vulnerability of AI and machine learning systems to adversarial tricking or spoofing amplifies the core problems of predictability and reliability.

11. Computer vision and image recognition are important applications of machine learning. These applications use deep neural networks (deep learning), of which the functioning is neither predictable nor explainable, and such networks can be subject to bias. More fundamentally, machines do not see like humans. They have no understanding of meaning or context, which means they make mistakes that a human never would.

12. It is significant that industry standards for civilian safety-critical autonomous robotic systems – such as industrial robots, aircraft autopilot systems and self-driving cars – set stringent requirements regarding: human supervision, intervention and deactivation – or fail-safe; predictability and reliability; and operational constraints. Leading developers of AI and machine learning have stressed the need to ensure human control and judgement in sensitive applications – and to address safety and bias – especially where applications can have serious consequences for people’s lives.

13. Civilian experience with autonomous systems reinforces and expands some of the ICRC’s viewpoints and concerns regarding autonomy in the critical functions of weapon systems. The consequences of using autonomous weapon systems are unpredictable because of uncertainty for the user regarding the specific target, and the timing and location of any resulting attack. These problems become more pronounced as the environment or the task become more complex, or freedom of action in time and space increases. Human-on-the-loop supervision, intervention and the ability to deactivate are absolute minimum requirements for countering this risk, but the system must be designed to allow for meaningful, timely, human intervention – and even that is no panacea.

14. All autonomous weapon systems will always display a degree of unpredictability stemming from their interaction with the environment. It might be possible to mitigate this

to some extent by imposing operational constraints on the task, the timeframe of operation, the scope of operation over an area and the environment. However, the use of software control based on AI – and especially machine learning, including applications in image recognition – brings with it the risk of inherent unpredictability, lack of explainability and bias. This heightens the ICRC's concerns regarding the consequences of using AI and machine learning to control the critical functions of weapon systems and raises questions about its use in decision-support systems for targeting.

15. This review of technical issues highlights the difficulty of exerting human control over autonomous (weapon) systems and shows how AI and machine learning could exacerbate this problem exponentially. Ultimately it confirms the need for States to work urgently to establish limits on autonomy in weapon systems.

16. It reinforces the ICRC's view that States should agree on the type and degree of human control required to ensure compliance with international law and to satisfy ethical concerns, while also underlining its doubts that autonomous weapon systems could be used in compliance with international humanitarian law in all but the narrowest of scenarios and the simplest of environments.