ICRC

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

Geneva, August 2019

CONTENTS

EXECUTIVE SUMMARY

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.

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.

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.

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.

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.

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.

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.

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**.

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.

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.

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.

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.

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.

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.

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.

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.

1. INTRODUCTION

New technological developments in autonomy, AI and robotics have broad applications across society, bringing both opportunities and risks. Military applications in armed conflict may bring benefits to the extent they help belligerents to minimize adverse consequences for civilians and ensure compliance with international humanitarian law. However, in weapon systems, they may also give rise to significant risks of unintended, and potentially unlawful, effects stemming from a lack of control. Indeed, the **ICRC's core concern with autonomous weapon systems is a loss of human control over the use of force**, which:

- has potentially serious consequences for protected persons in armed conflict
- raises significant legal questions regarding compliance with international humanitarian law
- prompts fundamental ethical concerns about human responsibility for life-and-death decisions.

States party to the Convention on Certain Convention Weapons have agreed that "human responsibility" for decisions on the use of weapon systems and the use of force "must be retained".[1] **The ICRC's view is that to retain human responsibility in this area States must now agree limits on autonomy in weapon systems by specifying the type and degree of human control required** to ensure compliance with international humanitarian law and other applicable international law, and to satisfy ethical concerns.[2]

The ICRC has published its views on the legal[3] and ethical[4] obligation to ensure human control and has proposed that key aspects include:

- human supervision, intervention and deactivation
- predictability and reliability
- operational constraints on tasks, targets, environments, time and space.[5]

In June 2018, the ICRC convened a round-table meeting with independent experts on autonomy, AI and robotics, to gain a better understanding of the technical aspects of human control, drawing on experience and lessons learned with civilian autonomous systems.[6] This report summarizes the discussions of that meeting and supplements them with additional research. It highlights key themes and conclusions from the perspective of the ICRC, and these do not necessarily reflect the views of the participants.

---

[1] United Nations, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, CCW/GGE.1/2018/3, 23 October 2018. Sections III.A.26(b) & III.C.28(f).

[2] ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019.

[3] *Ibid.* See also: ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 9–13 April & 27–31 August 2018. N. Davison, "Autonomous weapon systems under international humanitarian law", in United Nations Office for Disarmament Affairs, *Perspectives on Lethal Autonomous Weapon Systems*, United Nations Office for Disarmament Affairs (UNODA) Occasional Papers, No. 30, November 2017, pp. 5–18: https://www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law. ICRC, *Views of the ICRC on autonomous weapon systems*, 11 April 2016: https://www.icrc.org/en/document/views-icrc-autonomous-weapon-system.

[4] ICRC, *Ethics and Autonomous Weapon Systems: An Ethical Basis for Human Control?*, report of an expert meeting, 3 April 2018: https://www.icrc.org/en/document/ethics-and-autonomous-weapon-systems-ethical-basis-human-control.

[5] ICRC, *The Element of Human Control*, Working Paper, Convention on Certain Conventional Weapons (CCW) Meeting of High Contracting Parties, CCW/MSP/2018/WP.3, 20 November, 2018.

[6] The meeting, *Autonomy, artificial intelligence and robotics: Technical aspects of human control*, took place at the Humanitarium, International Committee of the Red Cross (ICRC), Geneva, on 7-8 June 2018. With thanks to the following experts for their participation: Chetan Arora, Subhashis Banerjee (Indian Institute of Technology Delhi, India); Raja Chatila Chatila (Institut des Systèmes Intelligents et de Robotique, France); Michael Fisher (University of Liverpool, United Kingdom); François Fleuret (École Polytechnique Fédérale de Lausanne (EPFL), Switzerland); Amandeep Singh Gill (Permanent Representative of India to the Conference on Disarmament, Geneva); Robert Hanson (Australian National University, Australia); Anja Kaspersen (United Nations Office for Disarmament Affairs, Geneva Branch); Sean Legassick (DeepMind, United Kingdom); Maite López-Sánchez (University of Barcelona, Spain); Yoshihiko Nakamura (University of Tokyo, Japan); Quang-Cuong Pham (Nangyang Technological University, Singapore); Ludovic Righetti (New York University, USA); and Kerstin Vignard (United Nations Institute for Disarmament Research, UNIDIR). The ICRC was represented by: Kathleen Lawand, Neil Davison, Netta Goussac and Lukas Hafner (Arms Unit, Legal Division); Laurent Gisel and Lukasz Olejnik (Thematic Unit, Legal Division); and Sasha Radin (Law and Policy Forum). Report prepared by Neil Davison.

2. AUTONOMY IN WEAPON SYSTEMS

2.1 Characteristics

The ICRC defines an autonomous weapon system as "**Any weapon system with autonomy in its critical functions**. That is, a weapon system that can **select** (i.e. search for or detect, identify, track, select) and **attack** (i.e. use force against, neutralize, damage or destroy**) targets without human intervention**." Autonomous weapon systems are not a discrete category of weapon, since autonomy in critical functions could be added to any weapon system.

These weapon systems self-initiate or trigger an attack or attacks in response to objects or persons detected in the environment, based on a general target profile. In other words, after initial activation or launch by a human operator, the weapon system – through its sensors, programming (software) and connected weapon(s) – takes on the targeting functions that would normally be carried out by humans. Consequently, the **user will not know the specific target, nor the exact timing and location of the attack that will result**. This means that autonomous weapon systems all introduce a degree of unpredictability into the consequences of the attack(s), creating risks for civilians and civilian objects and challenges for compliance with international humanitarian law. These weapon systems are clearly different from others – whether directly or remotely controlled by humans – where the user chooses the specific target, timing and location of the attack at the point of launch or activation (even if there may be a time-delay in reaching the target).

A weapon might have autonomy in its critical targeting functions without having "system-level" autonomy, i.e. autonomy in all other functions (such as flight or navigation). Furthermore, autonomy in critical functions is not dependent on technical sophistication; a weapon could be very simple and "unintelligent" in its design, but highly autonomous in its targeting functions. In other words, autonomous weapon systems do not necessarily incorporate AI and machine learning; existing weapons with autonomy in their critical functions generally use simple, rule-based control software to select and attack targets.[7]

2.2 Trends in existing weapons

A non-exhaustive study by the Stockholm International Peace Research Institute found 154 existing weapon systems with autonomy in some aspects of targeting, including 49 that fall within the ICRC's definition of autonomous weapon systems, and 50 that employ automatic target recognition as a decision-support tool for human operators, who then decide whether to authorize or initiate an attack.[8]

Existing autonomous weapon systems include:

- **air defence systems** – short and long range – with autonomous modes for shooting down incoming missiles, rockets, mortars, aircraft and drones
- **active protection systems**, which function in a similar way to protect tanks or armoured vehicles from incoming missiles or other projectiles
- some **loitering weapons** – a cross between a missile and a drone – which have autonomous modes enabling them to target radars based on a pre-programmed radio-frequency signature.

[7] ICRC, *ICRC Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019. Agenda item 5(b). This is one reason why the ICRC has seen notions of "automated" and "autonomous" weapons as interchangeable for the purpose of its legal analysis.

[8] V. Boulanin and M. Verbruggen, *Mapping the Development of Autonomy in Weapon Systems*, Stockholm International Peace Research Institute (SIPRI), November, 2017. ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, 2016, Report of an expert meeting: https://www.icrc.org/en/publication/4283-autonomous-weapons-systems; ICRC, *Autonomous weapon systems: Technical, military, legal and humanitarian aspects*, 2014, Report of an expert meeting: https://www.icrc.org/en/document/report-icrc-meeting-autonomous-weapon-systems-26-28-march-2014.

Generally, **current autonomous weapon systems are anti-materiel weapons** that employ relatively simple sensor and software technology to identify the signatures of pre-defined objects such as missiles, rockets, mortars, aircraft, drones, tanks, ships, submarines and radar systems. Almost all are **human-supervised in real time**; a human operator can intervene and divert or deactivate the system, and in many cases can verify a target before the attack takes place.

There are also significant **operational constraints** on:

- the types of **task** the weapons are used for – primarily the protection of ships, military bases or territory from incoming projectiles
- the **targets** they attack – only objects or vehicles
- the **environments** in which they are used – for example at sea or around military bases, where risks to civilians and civilian objects are lower than in populated areas
- the **timeframe** and **scope of operation** – autonomous modes are mostly activated for limited periods and the vast majority of systems are constrained in space and are not mobile.

There are **no autonomous weapon systems in use today that directly attack human targets** without human authorization. However, some countries have developed or acquired "sentry" weapons, which they deploy at borders and perimeters or mount on vehicles. These identify and select human targets autonomously but require human verification and authorization to fire.[9]

2.3 Possible future developments

Autonomy in targeting is a *function* that could be applied to any weapon system, in particular the rapidly expanding array of robotic weapon systems, in the air, on land and at sea – including swarms of small robots. This is an area of significant investment and emphasis for many armed forces, and the question is not so much whether we will see more weaponized robots, but whether and by what means they will remain under human control. Today's remote-controlled weapons could become tomorrow's autonomous weapons with just a software upgrade.

The **central element of any future autonomous weapon system will be the software**. Military powers are investing in AI for a wide range of applications[10] and significant efforts are already underway to harness developments in image, facial and behaviour recognition using AI and machine learning techniques for intelligence gathering and "automatic target recognition" to identify people, objects or patterns.[11] Although not all autonomous weapon systems incorporate AI and machine learning, this software could form the basis of future autonomous weapon systems. Software systems – whether AI-enabled or not – could directly activate a weapon, making it autonomous. However, early examples of

[9] Although manufacturers have offered versions with autonomous attack capability. See, for example: S. Parkin, "Killer robots: The soldiers that never sleep", *BBC*, 16 July 2015: http://www.bbc.com/future/story/20150715-killer-robots-the-soldiers-that-never-sleep.

[10] ICRC, *Artificial intelligence and machine learning in armed conflict: A human-centred approach*, 6 June 2019: https://www.icrc.org/en/document/artificial-intelligence-and-machine-learning-armed-conflict-human-centred-approach. P. Scharre, "Killer Apps: The Real Dangers of an AI Arms Race", *Foreign Affairs*, May/June 2019. S. Radin, "Expert views on the frontiers of artificial intelligence and conflict", *ICRC Humanitarian Law & Policy Blog*, 19 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict. M. Horowitz *et al.*, *Artificial Intelligence and International Security*, Center for a New American Security (CNAS), July 2018. R. Surber, *Artificial Intelligence: Autonomous Technology (AT), Lethal Autonomous Weapons Systems (LAWS) and Peace Time Threats*, ICT4Peace Foundation and the Zurich Hub for Ethics and Technology, 21 February 2018. D. Lewis, G. Blum, and N. Modirzadehm, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict (HLS PILAC), Harvard University, 31 August 2016.

[11] B. Schachter, *Automatic Target Recognition*, Third Edition, SPIE, 2018. SBIR, *Automatic Target Recognition of Personnel and Vehicles from an Unmanned Aerial System Using Learning Algorithms*, DoD 2018.1 SBIR Solicitation, 2018: https://www.sbir.gov/sbirsearch/detail/1413823.

AI and machine learning applications take the form of decision-support systems to "advise" human fighters on matters that include targeting decisions.[12]

Beyond software, other developments include shifts:

- from anti-materiel systems to anti-personnel systems
- from static (fixed) "defensive" systems to mobile "offensive" systems actively searching for targets over an area
- from single-platform systems to swarms of several hundred operating together[13]
- from use in armed conflict to use in law enforcement operations.[14]

3. HUMAN CONTROL

3.1 What is an autonomous system?

An autonomous (robotic) system or function is a closed loop ("sense-think-act"). The machine

- receives information from its environment through sensors ("sense")
- processes these data with control software ("think")
- based on its analysis, performs an action ("act") without further human intervention.

Autonomy, therefore, is the ability of the system to act without direct human intervention, although it is a continuum with various levels and many grey areas. In civilian robotics, some autonomous systems perform prescribed actions that are fixed in advance and do not change in response to the environment (such as an industrial manufacturing robot). These are sometimes referred to as "automatic". Other systems initiate or adjust their actions or performance based on feedback from the environment ("automated") and more sophisticated systems combine environmental feedback with the system's own analysis regarding its current situation ("autonomous"). Increasing autonomy is generally equated with greater adaptation to the environment and is sometimes presented as increased "intelligence" – or even "artificial intelligence" – for a particular task. That said, the perception of both autonomy and AI is constantly shifting, as advances in technology mean that some systems once considered "autonomous" and "intelligent" are now classed merely as "automated". Importantly, **there is no clear technical distinction between automated and autonomous systems**, nor is there universal agreement on the meaning of these terms, and for the remainder of this report we will use "autonomous" to represent both of these concepts of "systems that interact with their environment".

3.2 Human control over autonomous systems?

By definition, an autonomous system or function is to some degree out of human control. Nevertheless, humans can exert some control during design and development, at the point of activation for a specific task and during operation, for example by interrupting its functioning.[15] In the context of autonomous weapon systems, the International Panel on the Regulation of Autonomous Weapons

[12] See, for example: S. Freedberg Jr, "ATLAS: Killer Robot? No. Virtual Crewman? Yes." *Breaking Defense*, 4 March 2019: https://breakingdefense.com/2019/03/atlas-killer-robot-no-virtual-crewman-yes. D. Lewis, N. Modirzadeh, and G. Blum, "The Pentagon's New Algorithmic-Warfare Team", *Lawfare*, 2017: https://www.lawfareblog.com/pentagons-new-algorithmic-warfare-team. J. Keller, "DARPA TRACE program using advanced algorithms, embedded computing for radar target recognition", *Military & Aerospace Electronics*, 2015: http://www.militaryaerospace.com/articles/2015/07/hpec-radar-target-recognition.html.

[13] D. Hambling, *Change in the air: Disruptive Developments in Armed UAV Technology*, United Nations Institute for Disarmament Research (UNIDIR), 2018.

[14] M. Brehm, *Constraints and Requirements on the Use of Autonomous Weapon Systems Under International Humanitarian and Human Rights Law*, Geneva Academy of International Humanitarian Law and Human Rights, Academy briefing no. 9, May 2017, pp. 42–68.

[15] For an analysis of this concept applied to autonomous weapon systems see N. Davison, *op. cit.*

(iPRAW) has distinguished between "control by design" (i.e. in design and development) and "control in use" (i.e. in activation and operation), while stressing the importance of both.[16]

There is no universal model for optimal human-machine interaction with autonomous (robotic) systems, since the **need for human control, or the level of autonomy that one can tolerate, is linked to the complexity of the environment** in which the system operates **and the complexity of the task** it carries out. Generally, the greater the complexity in either the greater the need for direct human control and less tolerance of autonomy, especially for tasks and in environments where a system failure could kill or injure people or damage property, i.e. "safety-critical" tasks.[17] Use of an autonomous system in an uncontrolled, unpredictable environment carries a high risk of malfunctioning and unexpected results. Nevertheless, current technical developments in software – complex control software including but not limited to AI and machine learning techniques – seek to increase the level of autonomy that can be tolerated for more complex tasks in more complex environments.[18]

3.3 Modes of control

Human control over robotic systems can take several forms.

Direct control

Requires constant intervention by a human operator to directly or remotely control the functions of the system, which are therefore not autonomous.

Shared control

The human operator directly controls some functions while the machine controls other functions under the supervision of the operator. Examples include certain non-autonomous robotic weapon systems, such as armed drones. In these systems, a human operator directly (albeit remotely) controls the critical targeting functions, while the machine might control flight and navigation functions autonomously with human supervision.

Shared control aims to:

- exploit the benefits of human control (global situational awareness and judgement) and machine control (specific actions at high speed and accuracy)
- partly circumvent the limitations of human control (limited attention span and field of perception, stress and fatigue) and machine control (limited decision-making capacity, sensing uncertainties and limited situational awareness).

Supervisory control

A robotic system performs tasks autonomously while the human operator supervises, and the operator can provide instructions and/or intervene and take back control, as required.[19] In general, enabling a robotic system to perform tasks autonomously while retaining human supervisory control requires knowledge of how the system will function in the future – "predictive control" – so that the user can judge when intervention will be necessary, and in what form. This, in turn, requires knowledge of the

[16] iPRAW, *Concluding Report: Recommendations to the GGE*. International Panel on the Regulation of Autonomous Weapons (iPRAW), December 2018, p. 14.
[17] J. Knight, "Safety-critical Systems: Challenges and Directions", *Proceedings of the 24th International Conference on Software Engineering*, February 2002.
[18] However, many cutting-edge autonomous robotic systems, such as the humanoid and dog-like robots developed by Boston Dynamics, do not use AI and machine learning software.
[19] B. Siciliano and O. Khatib, (eds) *Springer Handbook of Robotics*, 2nd Edition, 2016, p. 1091. T. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*, MIT Press, 1992. For an analysis applying this concept to autonomous weapon systems see N. Sharkey, "Staying in the loop: Human supervisory control of weapons", in N. Bhuta *et al.* (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, 2016, pp. 23–38.

environment in the future; in other words a predictable environment. In the civilian world, supervisory control is often used in applications where direct or shared control of the robotic system is not possible due to communication delays between instructions sent by the operator and the subsequent action of the system, such as in systems operating in outer space or deep under the sea. Most existing autonomous weapon systems operate under some form of supervisory control for specific tasks in highly constrained – and therefore relatively predictable – environments.[20]

3.4 Human-on-the-loop

In most real-world situations, the operating environment is dynamic and unpredictable, and predictive control is therefore difficult. However, **human supervisory control enables operators to exert some control through "human-on-the-loop" supervision and intervention**. There may be more than one loop through which the human can intervene, with different results, such as a low-level control loop for specific functions (control level) and/or a high-level control loop for more generalized goals (planning level).

In any case, effective human-on-the-loop supervision and intervention require the human operator to have continuous **situational awareness**, **enough time to intervene** (i.e. override, deactivate or take back control) and a **mechanism through which to intervene**, notably a permanent communication link (for remotely operating systems) and/or direct physical controls, that enable the user to take back control or deactivate the system.

Unfortunately, the human-on-the-loop model – even if it satisfies the above criteria – is not a magic bullet for ensuring effective control over autonomous (robotic) systems because of well-known **human-machine interaction problems**, in particular:

- **automation bias** – or over-trust in the machine – where humans place too much confidence in the operation of an autonomous machine
- **lack of operator situational awareness** (insufficient knowledge of the state of the system at the time of intervention, as explained below)
- the **moral buffer**, where the human operator shifts moral responsibility and accountability to the machine as a perceived legitimate authority.[21]

It is also necessary to consider whether "safe takeover" is possible, and at which point in time. There may be negative consequences if there is limited time available – due to the speed of the robotic operation – and/or a delay before the human operator can take back control. One example would be a human operator not having time to take back control over a self-driving car to apply the brakes and prevent a collision. This type of problem already arises with existing human-supervised autonomous weapon systems, such as air defence systems, which have shot down aircraft in "friendly fire" accidents before an operator could deactivate them.[22] Complicating matters, immediate interruption of an autonomous system by a human operator can sometimes be more dangerous than waiting to intervene. An aircraft cannot stop in mid-air, for example, and a switch from autopilot to manual control can be catastrophic if the pilot does not have current situational awareness.[23] In sum, **one cannot assume that human-on-the-loop intervention will be an effective way of mitigating the risks of loss of control inherent to autonomous (robotic) systems**.

A human operator **override function** – effectively a "big red button" to deactivate the system – is generally part of the design of civilian autonomous (robotic) systems that perform safety-critical tasks.

[20] ICRC, *Autonomous weapon systems: Implications of increasing autonomy in the critical functions of weapons*, *op. cit.*
[21] M Cummings, "Automation and Accountability in Decision Support System Interface Design", *Journal of Technology Studies*, Vol. XXXII, No. 1, 2006.
[22] J. Hawley, *Automation and the Patriot Air and Missile Defense System*, Center for a New American Security (CNAS), 25 January 2017.
[23] R. Charette, "Air France Flight 447 Crash Causes in Part Point to Automation Paradox", *IEEE Spectrum*, 10 July 2012: https://spectrum.ieee.org/riskfactor/aerospace/aviation/air-france-flight-447-crash-caused-by-a-combination-of-factors.

This can help avert negative outcomes, but not always, given the problems of human-machine interaction and safe takeover. **Built-in fail-safe mechanisms** are therefore an important way of avoiding negative consequences in situations where human intervention is neither possible nor safe. It is possible to design a fail-safe mechanism to deactivate the system in specific circumstances, such as when it encounters an unknown environment, or when a malfunction occurs. However, even a fail-safe such as an immediate stop can have negative consequences, depending on the nature of the system and the environment, such as self-driving car travelling at speed on a busy highway.

One lesson from civilian robotics is that the appropriate form of human control may depend on the specific task the system is carrying out, the environment of use and, in particular, the timescale over which it operates. In weapon systems, **maintaining either direct (human) control over critical functions of targeting or shared control** – where critical functions remain under direct control while other functions may be autonomous – **is the most effective way of addressing the unpredictability caused by autonomy in targeting** (see also Sections 2 and 4). Where these critical functions are autonomous, supervisory control with a **"human-on-the-loop" may only be meaningful and effective if there is enough time for the operator** to select and approve one of several options proposed by the system, to override and take back control, or to deactivate the system, before the weapon fires at a target. Given the importance of the time available for effective human intervention, one approach to human control over autonomous weapon systems might be to **design safeguards that ensure there is always an alert for the operator, and enough time available for human intervention or authorization**, before the system initiates an attack.

4. PREDICTABILITY AND RELIABILITY

Predictability and reliability are at the heart of discussions about autonomy in weapon systems, since they are essential to ensuring compliance with international humanitarian law and avoiding adverse consequences for civilians. They are also essential for military command and control.[24] It is important to be clear what we mean by predictability and reliability, as these terms are sometimes used and understood in different ways.

Predictability

In discussions about autonomous weapon systems, the ICRC has understood predictability as the ability to "say or estimate that a specified thing will happen in the future or will be a consequence of something", in other words **knowledge of how the weapon system will function in any given circumstances of use**, including the effects that will result.[25] This is **predictability in a broad sense of knowing the *outcome*** that will result from activating the autonomous weapon system in a particular circumstance. A sub-component of this is **predictability in a narrow sense of knowing the *process*** by which the system functions and carries out a specific task or function. Both are important for ensuring compliance with international humanitarian law.

Reliability

Reliability is "the quality of being trustworthy or performing consistently well", in other words **how consistently the weapon system will function as intended**, without failures (malfunctions) or unintended effects. Reliability is, in effect, a measure of how often a system fails, and is a narrower concept than predictability. A given system can reliably carry out a specific function without being predictable in the effects that will result in a particular circumstance. A system may be predictable in its general functioning and likely effects, but subject to frequent failures.

[24] ICRC, *ICRC Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019. Agenda item 5(c).
[25] N. Davison, *op. cit.*

Examples

Anti-personnel landmines

Mines, which have been described as "rudimentary autonomous weapon systems"[26], illustrate the differences between reliability and predictability, and between broad and narrow notions of predictability, as well as the role of the environment in unpredictability. An anti-personnel landmine might be highly reliable (i.e. always detonates when activated by a certain weight) and highly predictable in a narrow sense (i.e. triggers when anything over a certain weight presses on it). Despite this, landmines are highly *unpredictable* in a broad sense of the consequences of their use, because it is not known who (or what) will trigger them, or when. This type of unpredictability has led to indiscriminate effects in most contexts where anti-personnel mines have been used, with severe consequences for civilians, and led to the prohibition of anti-personnel landmines in 1997 through the Anti-Personnel Mine Ban Convention.

Anti-radar loitering weapons

Anti-radar loitering weapons in autonomous mode illustrate the same issue. A loitering weapon might be very reliable (i.e. always detects a radar signature and then moves towards it and detonates) and highly predictable in a narrow sense (i.e. only attacks when it detects a specific type of radar signature). And yet it remains highly unpredictable in a broad sense of the consequences of an attack, because the user does not know which radar it will attack, where the attack will take place or when, or whether there are civilians or civilian objects near the target.

As these examples illustrate, **autonomous weapon systems are unpredictable in a broad sense**, because they are triggered by their environment at a time and place unknown to the user who activates them. Moreover, developments in the **complexity of software control systems – especially those employing AI and machine learning – may add unpredictability in a narrow sense** of the process by which the system functions (see Section 5). Unpredictability raises questions regarding compliance with international humanitarian law since it will be difficult for a commander or operator to comply with their legal obligations regarding the conduct of hostilities if they cannot foresee the consequences of activating an autonomous weapon system.[27]

Factors affecting predictability and reliability

Predictability and reliability are not inherent properties of the technical design of an autonomous robotic system. They also depend on:

- the nature of the environment
- the interaction of the system with the environment
- the complexity of the task.

An autonomous robotic system that functions predictably in a specific environment may become unpredictable if that environment changes or if it is used in a different environment. Predictability and reliability in carrying out a task will also depend on the complexity of the task and the options available to the system, which will constrain its eventual action (output) in a given situation.

[26] United States Department of Defense, *Department of Defense Law of War Manual*, Section 6.5.9.1, Description and Examples of the Use of Autonomy in Weapon Systems, 2015, p. 328: "Some weapons may have autonomous functions. For example, mines may be regarded as rudimentary autonomous weapons because they are designed to explode by the presence, proximity, or contact of a person or vehicle, rather than by the decision of the operator."

[27] ICRC, *ICRC Statement to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019. Agenda item 5(a).

The technical design of the system will also have a significant impact. **Increased complexity**, including in the software and the sensors that collect data – for example, combining multiple sensor inputs and/or increasing the complexity of the algorithm used to analyse input data – **will lead to less predictability, raising specific concerns about accidents and reliability**.[28] This is the case even for deterministic (rule-based) software and is even more applicable to AI and machine learning approaches, which may be unpredictable by design (see Section 5). Even deterministic systems do not function in a broadly predictable fashion, owing to complexity (in design and task) and interaction with a varying environment. **Swarming robots would raise particularly serious concerns regarding unpredictability**, since the interaction of multiple systems represents an immense increase in complexity, which can also lead to "emergent" unpredictable behaviours.[29]

Reducing unpredictability

Setting boundaries on the operation of an autonomous robotic system is one approach to reducing unpredictability. One way of achieving this is to **constrain the environment** in which the system operates. Although there will always be unknown environmental variables in the real world, some environments – such as airspace and undersea – are generally less complex and therefore less challenging in terms of predicting the environment's impact on how a system will function; the less the complexity and variation in the environment, the higher the potential level of predictability. This is one reason why it is much easier to ensure the predictability and reliability of autopilot systems for aircraft than it is for self-driving cars. Additional **constraints on the timeframe of autonomous operation and scope of operation over an area** can also reduce unpredictability by limiting the exposure of the system to variations over time in the environment in which it is operating. These are all factors that one may need to consider in discussions about ensuring human control over weapon systems.

4.1 Testing

Verifying and validating autonomous robotic systems that respond or adapt to their environment, in order to ensure sufficient predictability and reliability, brings its own challenges. Testing normally includes computer simulations and real-world physical tests to assess the response of the system in the different circumstances it may encounter. However, **it is not possible to test all the potential inputs and outputs of the system for all circumstances**, or even to know what percentage of the possible outputs one has tested. This means that it is difficult to formally verify and validate the predictability of the system and its reliability, or probability of failure. Considering weapon systems, it is therefore difficult to ensure that an autonomous weapon system is capable of being used in compliance with international humanitarian law,[30] especially if the system incorporates AI – and particularly machine learning – control software.[31]

The more complex the environment, the more acute the problem of verification and validation. Given the limits of testing in the real world, computer simulations are used to increase the number of scenarios that can be tested. However, simulations bring their own difficulties, as building an accurate simulation is difficult and requires knowledge of all critical scenarios and how to re-create them faithfully. Simulations cannot generally replicate the real-world environment, even for simple tasks. The

[28] UNIDIR, *Safety, Unintentional Risk and Accidents in the Weaponization of Increasingly Autonomous Technologies*, UNIDIR, 2016.
P. Scharre, *Autonomous Weapons and Operational Risk*, CNAS, February 2016.
[29] P. Scharre, *Robotics on the Battlefield Part II: The Coming Swarm*, CNAS, October 2014.
[30] ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts,* report for the 32nd International Conference of the Red Cross and Red Crescent, Geneva, October 2015, pp. 38–47: https://www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts.
[31] N. Goussac, "Safety net or tangled web: Legal reviews of AI in weapons and war-fighting", *ICRC Humanitarian Law & Policy Blog*, 18 April 2019: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting. D. Lewis, "Legal reviews of weapons, means and methods of warfare involving artificial intelligence: 16 elements to consider", *ICRC Humanitarian Law & Policy Blog*, 21 March 2019: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider.

design of a simulation can also introduce bias in the testing results and in the functioning of AI algorithms trained using simulations before being deployed (see Section 5). The question of how to test the accuracy of a simulation can therefore become an indefinite problem.

An example – self-driving cars

Testing in real traffic conditions is used to assess the reliability and predictability of self-driving cars, but it is very hard to test for "edge cases" –scenarios that occur relatively rarely but might result in failure of the system or unpredictable consequences. Obtaining enough data may require millions or billions of kilometres of testing. Furthermore, even if one combines real-world tests and simulations, it is impossible to test for every possible scenario.

This being so, any assessment of the predictability and reliability of an autonomous robotic system can only ever be an estimate, and it is difficult to provide a guarantee of performance – one can speak only in terms of probability. **Quantifying the predictability and reliability of an autonomous system – or function – is therefore difficult**, and it may be hard to decide what level would be sufficient. For example, if the sensors and image-recognition system for a self-driving car identify an object as "89% stop sign" or "94% pedestrian", what does this mean in terms of predictability and reliability? Must these figures be 99.9%? And if so, how can one be certain of having achieved this figure if it is only ever an estimate? Stringent standards exist for simpler autonomous systems – such as aircraft autopilots – (see Section 7.1), but these methods do not yet extend to more complex systems, such as self-driving cars. The implications for any use of these technologies in weapon systems are clearly significant.

An additional complication – adversarial conditions

Adversarial conditions bring further complications in testing – and in the real world (see also Section 6). By "adversarial conditions" we mean changes to the environment designed to trick or spoof the system. A well-known example is research showing that it is possible to trick the image recognition systems used in self-driving cars into thinking a stop sign is a speed limit sign just by placing small stickers on the sign.[32] This is already a significant problem in environments, such as city streets, where one might expect that most people are not deliberately trying to fool the system.[33] However, in the context of armed conflict, and weapon systems with autonomous functions, the problem would be exponentially worse, as the user could assume their adversary would constantly, and deliberately, be attempting to spoof these systems.[34]

5. ALGORITHMS, AI AND MACHINE LEARNING

An algorithm is a sequence of programming instructions – or rules – which, when executed by a computer, performs a calculation or solves a given problem.[35] A computer using an algorithm has the advantage, compared to humans, that it can process large amounts of data very quickly and accurately.

In general, these **deterministic (rule-based) algorithms are predictable** in their output for a given input. **But this does not mean they are necessarily predictable in the consequences of applying that output** in a given circumstance (see Section 4 on narrow versus broad notions of predictability). Deterministic algorithms are also relatively transparent in their programming, and therefore understandable (assuming one has access to the source code). In rule-based programming, the functioning of the algorithm (potential inputs and the resulting outputs of the system they control) is fixed

[32] K. Eykholt, *et al.*, *Robust Physical-World Attacks on Deep Learning Models*, Cornell University, v.5, 10 April 2018: https://arxiv.org/abs/1707.08945.
[33] R. Brooks, "The Big Problem With Self-Driving Cars Is People", *IEEE Spectrum*, 27 July 2017: https://spectrum.ieee.org/transportation/self-driving/the-big-problem-with-selfdriving-cars-is-people.
[34] P. Scharre, 2016, *op. cit.*
[35] Oxford English Dictionary, *Algorithm*: https://en.oxforddictionaries.com/definition/algorithm.

at the point of design. Such algorithms enable an autonomous robotic system to react to its environment or be triggered by it, but the system has little ability to adapt to that environment. This is how many existing autonomous weapon systems function, such as air defence systems; sensors detect an incoming object and the algorithm controlling the system triggers the weapon to fire if the object is moving within a certain range of speeds and within a specific trajectory.

Adaptability to the environment

The more complex the environment, the greater the adaptability needed to ensure the autonomous functioning of a robotic system. For example, an autonomous robot navigating on land will require greater adaptability than one navigating in the air or under the sea. One way to increase adaptability is to use more complex algorithms, which can assess the environment and act to adapt to that environment in carrying out a task. A system is given a generalized goal or objective, and the algorithm decides how to achieve it. For example, a mobile autonomous robotic system is given the goal of moving to a specific destination but the system itself determines the route it will take based on its programming and on data inputs from sensors that detect its environment. The user may also provide it with other sources of data at the outset, such as a map of the area in the case of a self-driving car. This contrasts with a more directive algorithm, which would specify both the destination and the route and therefore not allow the system to adapt to its environment.

Increasing **adaptability in an autonomous system is generally equated with increasingly "intelligent" behaviour – or AI**. Definitions of AI vary, but they are computer programs that carry out tasks – often associated with human intelligence – that require cognition, planning, reasoning or learning. What is considered AI has changed over time: and autonomous systems once considered "intelligent" – such as aircraft autopilot systems – are now seen as merely automated.[36] There is growing interest in the military application of AI for purposes that include weapon systems and decision support more broadly, whether for targeting or for other military applications.[37]

5.1 Machine learning

Rule-based AI systems – "expert systems" – are used in autonomous robotic systems that can perform increasingly complex tasks without human intervention, such as robots that can walk and move and the overall control software for self-driving cars. However, there is a significant focus today on a particular type of AI: machine learning.

What is machine learning?

Machine learning systems are **AI systems that are trained on – and learn from – data, which define the way they function**.

Instead of following pre-programmed rules, machine learning systems build their own model (or "knowledge") based on sample data input representing the input or task they are to learn, and then use this model to produce their output, which may consist of carrying out actions, identifying patterns or making predictions.[38]

Unlike when they are developing other AI algorithms such as expert systems, described above, developers do not specify how the algorithm functions with rules, or provide it with knowledge about the task or the environment; the functioning of a machine learning system is data driven. The outputs of

[36] UNIDIR, *Artificial Intelligence, a primer for CCW delegates*. The Weaponization of Increasingly Autonomous Technologies, UNIDIR Resources No. 8, p. 2.
[37] See for example S. Hill, and N. Marsan, "Artificial Intelligence and Accountability: A Multinational Legal Perspective", in *Big Data and Artificial Intelligence for Military Decision Making*, Meeting proceedings STO-MP-IST-160, NATO, 2018.
[38] For a useful overview of machine learning see Royal Society, *Machine learning: the power and promise of computers that learn by example*, April 2017, pp. 16–31.

these systems depend on the type of learning process and the resulting model that the system learns, which in turn depends on the data to which the algorithm is exposed. As a result, the **outputs of machine learning systems and the functions they control are much more unpredictable** than those of expert systems encoded with pre-defined instructions and knowledge, since the developer or user does not know what the system learns.[39]

Approaches to machine learning

There is a wide variety of machine learning approaches, which differ in the way learning takes place, the nature of the models and the problems they solve or the tasks they perform. However, they generally follow a two-step process. **First, there is a training phase** during which data are provided by the developer as inputs from which the algorithm will develop its model (or knowledge) as the output.

This training may take the form of:

- **supervised learning** – where developers categorize or label the data inputs (e.g. the content of an image)

or

- **unsupervised learning**, where the algorithm creates its own categories based on the training data (e.g. unlabelled images).

The **second phase is the deployment of the algorithm**, where it performs a task or solves a problem. Here the algorithm is exposed to data in the environment as inputs for that task and computes a solution, recommendation or prediction using the model it developed during the training phase.

These two steps are usually kept separate in most of today's civilian applications, and training stops before the algorithm is deployed (**off-line learning**), as combining these stages leads to increased errors and failure. However, some algorithms continue learning after deployment (**online learning**), thereby constantly changing the model on which they process data inputs to produce their results, or outputs. This adds an additional layer of complexity and unpredictability owing to changes in functioning in response to real-time data. One example of this was the conversational chat bot that was quickly reduced to expressing extremist views.[40]

One general difficulty with training machine learning algorithms is that it is hard to know when training is complete, i.e. when the algorithm has acquired a model that is sufficiently good for it to solve a problem based on data it is exposed to in the environment during that task. Furthermore, one can only assess the performance and reliability of the system for a given task against the testing and validation data set, since it is not possible to train an algorithm with every possible data input it might encounter in the environment.

Machine learning techniques

We can divide machine learning techniques into those where there is some organized structure to capture knowledge in a model and those where there is no such structure, such as a neural network.

It is possible for a user to interrogate a machine learning system that structures the knowledge it has learned, to try and understand why the algorithm has produced a certain output, although the complexity and quantity of the information available can make this very difficult.

Unstructured machine learning systems, on the other hand, produce their output without any explanation. They **constitute "black boxes", in that we do not know how or why they have produced a**

[39] M. Lopez-Sanchez, "Some Insights on Artificial Intelligence Autonomy in Military Technologies", in *Autonomy in Future Military and Security Technologies: Implications for Law, Peace, and Conflict*, The Richardson Institute, Lancaster University, UK, 10 November 2017, pp. 5–17.

[40] D. Alba, "It's Your Fault Microsoft's Teen AI Turned Into Such a Jerk", *Wired*, 25 March 2016, https://www.wired.com/2016/03/fault-microsofts-teen-ai-turned-jerk.

given output. Efforts to peer into the black box using another algorithm are at an early stage of development, the aim of this "explainable AI" being to provide the user with an explanation as to why a machine learning system has produced a particular output.

What are machine learning systems used for?

A machine learning algorithm can tackle a range of problems and tasks, either as part of a pure software system or when controlling a physical robot. Tasks include:

- **classification** (to which category does the data belong?)
- **regression** (how does input data relate to the output?)
- **clustering** (which data inputs are similar to each other?).[41]

Uses for classification include image recognition. In this instance, the machine learning algorithm is trained (supervised or unsupervised) with data in the form of images, such as cats and dogs. Once training is complete, the algorithm analyses new data, classifying images according to categories (e.g. cat, dog or neither). Most current **image recognition applications employ deep neural network (deep learning) techniques**, which produce a classification (output) without any explanation as to how or why they placed an image in a particular category (see Section 6.)

5.1.1 Reinforcement learning

Reinforcement learning differs from supervised and unsupervised learning in that the algorithm is not given a specific training data set to build its model. In training, the **algorithm uses experience acquired through interactions with the environment to learn how to carry out a specific task**. The developer gives the algorithm a **goal**, or **"reward function"** (e.g. to win a game) and the algorithm then builds a model (a strategy to win the game in this example) based on trial-and-error interaction with the training environment. Then, in deployment, the algorithm uses this model to solve the problem (i.e. play and win the game). The algorithm is designed – or rather designs itself – based on this goal, rather than on specific training data.

Examples of reinforcement learning include learning a game and then beating human competitors (e.g. Deep Mind's AlphaGo),[42] and its main area of application is in decision-making and strategy, as opposed to systems that build relationships between input data and outputs, such as image recognition. Reinforcement learning is also being used to develop robots that might be able to explore unknown environments, albeit with limited success to date, especially for complex tasks. Current robotic systems deployed in the real world, such as self-driving cars, generally use traditional rules-based AI methods for decision-making and control aspects, and machine learning for computer vision and image processing.

Risks with reinforcement learning

While reinforcement learning offers new capabilities, it also brings risks, especially if used for safety-critical tasks. While the human developer defines the goal (reward function) and can exert some control over the training environment (which is usually a simulation), **the way in which the algorithm will learn, and then perform a task after deployment, is entirely unpredictable and often leads to unforeseen solutions to the task**.[43] One way to understand this is to think of a drop of water landing on the top of a mountain, the structure of which is completely unknown to you. You could predict, based on general understanding of gravity and the fact that mountains are elevated, that the drop of water will end up in the lake below. But you cannot know what route it will take, what will happen along the way, or when it will arrive, nor could you retrace its journey to understand how it arrived.

[41] Royal Society, *op. cit.*, p. 31.
[42] D. Silver, *et al.*, "Mastering the game of Go without human knowledge", *Nature*, Vol. 550, 19 October 2017, pp. 354–359.
[43] J. Lehman, *et al.*, *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities,* Cornell University, v.3. 14 August 2018: https://arxiv.org/abs/1803.03453.

Another major difficulty is transferring the results of reinforcement learning from computer simulations to robotic systems in the real world, known as the "sim-to-real" problem.

Specifying an appropriate reward function can be difficult, and even more so for complex tasks, because the way the goal is formulated can cause the algorithm to learn and perform in an unforeseen way. For example, a simulated six-legged robot that was given the reward function of walking with minimum contact between its feet and the ground learned to flip over and "walk" on its back using its elbows, achieving zero contact between its feet and the ground.[44]

Additional problems include:

- **preventing human override** – the algorithm may learn to prevent a human user from deactivating it (the "big red button problem");[45]
- **reward hacking** or gaming – the algorithm learns to exploit errors in the goal it has been given, leading to unpredictable consequences;[46]
- **emerging behaviours** – the algorithm carries out actions unrelated to the main goal.[47]

All these difficulties become even more acute when two or more reinforcement learning systems interact with each other, leading to extreme complexity and unpredictability. In theory, a reinforcement learning system might even learn to set or adjust its own goal, but such concerns are speculative as far as current technologies are understood; they may well perform a task in unpredictable ways but will not suddenly undertake a completely different task.

Machine-learning systems are also particularly vulnerable to "adversarial conditions" – changes to the environment designed to fool the system, or the use of another machine-learning system to produce adversarial inputs or conditions using a generative adversarial network (see also Section 6).

5.2 Trust in AI

Trust in AI and autonomous systems is a major area of enquiry, especially as regards their use for safety-critical applications or where they have other implications for human life and personal freedom.[48] Some have raised concerns about assumptions of the accuracy of analyses, or predictions, made by machine learning systems that are trained on past, limited, data sets. For example, the way many systems are developed means that assessments of their accuracy assume that the training data provides a correct representation of any data the algorithm may encounter "in the wild" during a task, whereas this may not be the case.

There are also concerns regarding the "**bias-variance trade-off**":

- **bias** in an algorithm makes it too simple, preventing it from identifying key patterns in new data ("underfitting")
- **variance** makes the algorithm too sensitive to the specific data it was trained on ("overfitting")[49], which means it cannot generalize its analysis when exposed to new data.

Improving bias can worsen variance and vice-versa.[50]

---

[44] J. Lehman, *et al.*, *op. cit.,* pp. 13–14.
[45] Orseau, L. and Armstrong, S., *Safely Interruptible Agents*, DeepMind, 1 January 2016: https://deepmind.com/research/publications/safely-interruptible-agents.
[46] D. Amodei, *et al.*, *Concrete Problems in AI Safety*, Cornell University, v.2, 25 July 2016: https://arxiv.org/abs/1606.06565.
[47] J. Leike, *et al.*, *AI Safety Gridworlds*, Cornell University, v.2, 28 November 2017, https://arxiv.org/abs/1711.09883.
[48] The Partnership on AI*, Safety-Critical AI: Charter*, 2018: http://www.partnershiponai.org/wp-content/uploads/2018/07/Safety-Critical-AI_-Charter.pdf.
[49] Oxford English Dictionary, *Overfitting*: https://en.oxforddictionaries.com/definition/overfitting.
[50] S. Geman, E. Bienenstock and R. Doursat, "Neural networks and the bias/variance dilemma", *Neural Computation*, Vol. 4 No. 1, 1992, pp. 1–58.

5.2.1 Bias

Bias in AI and machine learning algorithms is a core problem that can have many facets.[51] Types of bias include the following:

Training data bias

Perhaps the most common form of bias. Since machine learning algorithms learn using training data to refine their models, limits on the quantity, quality and nature of this data can introduce bias into the functioning of the algorithm.

Algorithmic focus bias

The algorithm gives different – or inappropriate – weighting to different elements of the training data and/or ignores some aspects of the data, leading, for example, to conclusions that are not supported by the data.

Algorithmic processing bias

The algorithm itself introduces bias in the way it processes data. Developers often introduce this type of bias or "regularization" intentionally as a way of counteracting other biases – for example to limit the problem of overfitting, or to account for limitations in the training data set.

Emergent bias

Emergent bias can cause an algorithm to function in unexpected ways owing to feedback from the environment. It is related to the context in which an algorithm is used, rather than to its technical design or the training data.[52]

Transfer context bias

An algorithm is used outside the context in which it was designed to function, possibly causing it to fail or behave unpredictably.

Interpretation bias

A user (human or machine) misinterprets the output of the algorithm, especially where there is a mismatch between information provided by the system and the information that the user requires to take a particular decision or perform a task.

5.2.2 Explainability

One way to build trust in an AI algorithm and its output is to provide explanations for how it produced its output that the user can interpret. One can then use these explanations to fine tune the model that the algorithm uses to produce its output, and thereby to address bias. However, **"explainability" is a fundamental problem for machine learning algorithms** that are not transparent in the way they function and provide no explanation for why they produce a given output (see Section 5.1).

Even when explanations are available, the question remains of whether one can extrapolate the trust built by analysing specific training data to trust in analysis of a general data set after deployment. Building trust in the model is more difficult, because the number of potential inputs in the environment may be infinite. This is currently a concern with self-driving cars; even after billions of kilometres of

[51] UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, UNIDIR Resources No. 9. D. Danks, and A. London, "Algorithmic Bias in Autonomous Systems", *Twenty-Sixth International Joint Conference on Artificial Intelligence*, August 2017.

[52] B. Friedman, and H. Nissenbaum, "Bias in computer systems", *ACM Transactions on Information Systems*, Vol. 14 No. 3, July 1996, pp. 330–347.

testing in simulations, and millions of kilometres of testing in real-world driving situations, it is hard to know when one has carried out enough testing and how the system will respond in unpredictable circumstances, to be confident that the system can be safely deployed.

A key issue is that **algorithmic bias remains a problem even with a human on the loop** to oversee the operation of an algorithm and approve the taking of certain actions based on its output, such as when a system advises a human decision-maker (decision support). Examples include systems that advise doctors on diagnoses or judges on sentencing.[53]

5.3 Implications for AI and machine learning in armed conflict

For the reasons of unpredictability, **most current civilian applications of machine learning do not perform safety-critical tasks, or if they do, they retain a human on the loop** to decide on or authorize specific actions. Linking this analysis to considerations of autonomous weapon systems, it seems that **AI – and especially machine learning – would bring a new dimension of inherent unpredictability by design**, which raises doubts as to whether they could ever lawfully be used to control the critical functions of selecting and attacking targets. These factors, together with issues of bias and lack of explainability, also raise concerns about the use of machine learning in decision-support systems for targeting and for other decisions in armed conflict that have significant consequences for human life. As well as technical issues, there are important questions about how to ensure a human-centred approach to the use of AI that maintains human control and judgement.[54]

6. COMPUTER VISION AND IMAGE RECOGNITION

Computer vision is a **major application of machine learning systems**, analysing digital images, video and the world around us.

These systems perform a variety of tasks, including:

- image classification (describing an image as a whole)
- object recognition (identifying specific objects within an image)
- scene understanding (describing what is happening in an image)
- facial recognition (identifying individual faces, or types of feature)
- gait recognition (identifying a person by the way they walk)
- pose estimation (determining the position of a human body)
- tracking a moving object (in a video)
- behaviour recognition (determining emotional states and behaviours using "affective computing").

Prominent civilian applications include self-driving cars, medical image processing (for example to aid doctors with diagnoses) and surveillance systems in law enforcement. However, parties to conflicts also use computer vision, for surveillance and intelligence analysis purposes such as identifying objects in video feeds from drones,[55] and it is being developed for automatic target recognition.[56]

Most **computer vision algorithms use deep convolutional neural networks, which means they cannot provide an explanation for their analysis**, and the sheer quantitative complexity makes it difficult to

[53] AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.
[54] See ICRC, *Artificial intelligence and machine learning in armed conflict: A human-centred approach*, *op. cit.*
[55] D. Lewis, N. Modirzadeh and G. Blum, 2017, *op. cit.*
[56] R. Hammoud and T. Overman, "Automatic Target Recognition XXIX", *Proceedings Vol. 10988, SPIE Defense & Commercial Sensing, 14–19 April 2019*: https://www.spiedigitallibrary.org/conference-proceedings-of-spie/10988.toc. B. Schachter, (2018) *op. cit.*

predict or understand how they produce their output (see Section 5.1). Furthermore, their performance is largely determined by the quality and completeness of the training data, and consequently even with large data sets these systems are likely to exhibit training data bias (see Section 5.2).

The semantic gap

But there is a more **fundamental problem with the use of computer vision** to replace or supplement human vision in analysing the world around us: the "semantic gap". What we mean by this is that **humans and machines "see" very differently**.[57] For a computer vision algorithm, an object in an image – a cat for example – is represented by a large three-dimensional series of numbers corresponding to pixels in the image. After having been trained on images of cats, the algorithm may be able to identify a cat in a particular image. However, unlike humans, the **algorithm has no understanding of the meaning or concept** of a cat (i.e. a mostly domesticated carnivorous mammal with highly developed hearing that hunts at night and sleeps most of the day). This lack of understanding means **algorithms can make mistakes that a human never would**, such as classifying a cat in an image as a football.

Algorithms can learn to make basic associations between an object and its context (e.g. "cat on a chair") but this still does not imply an understanding of the context. These associations can give a misleading sense of the algorithm's capability and can lead to inaccurate results: for example, an image classification algorithm trained on images of cats on chairs might only identify cats when they are on chairs, or may classify an image containing a chair as a cat. These are also mistakes that a human would never make. It is **not difficult to imagine the serious consequences if an image recognition system in a weapon system were to make this kind of mistake**.

Claims that "machines can now see better than humans" do not tell the full story, and humans and machines carry out tasks differently. Computer vision algorithms may be able to classify objects in a set of test images into specific categories more quickly and accurately than a human can,[58] but while effectiveness in carrying out this task is valuable, the fact that an algorithm cannot understand the meaning of the objects remains a problem. This core difference – and the mistakes that can occur – highlight the risks of using such systems for safety-critical tasks. This partly explains why **most civilian applications of image recognition that have consequences for human safety** – such as diagnosing skin cancer – **are used to advise human decision-makers rather than replace them**.[59] For example, the system can help identify a melanoma, but decisions on diagnosis and treatment are made by the doctor, who has the benefit of contextual understanding and judgement, together with information from other sources (such as patient history and physical examinations). As regards applications in weapon systems, this is probably why armed forces currently use such systems to automate the analysis of images and video, but not to act on this analysis and initiate an attack or take other decisions that could have serious consequences.

Reliability

For an algorithm to be useful in a real-world application, developers need to minimize the number of false positives, i.e. cases in which the algorithm incorrectly identifies an object. However, reducing this sensitivity can also lead to the algorithm missing objects that it should have identified – false negatives.

---

[57] A. Smeulders, *et al.*, "Content-Based Image Retrieval at the End of the Early Years", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 22, No. 12, December 2000, pp. 1349–1380.

[58] A. Hern, "Computers now better than humans at recognising and sorting images", *Guardian*, 13 May 2015: https://www.theguardian.com/global/2015/may/13/baidu-minwa-supercomputer-better-than-humans-recognising-images.
O. Russakovsky, *ImageNet Large Scale Visual Recognition Challenge*, Cornell University, v.3, 30 January 2015: https://arxiv.org/abs/1409.0575.

[59] H. Haenssle *et al.*, "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists", *Annals of Oncology*, Vol. 29, No. 8, August 2018, pp. 1836–1842. A. Trafton, "Doctors rely on more than just data for medical decision making", *MIT News Office*, 20 July 2018: http://news.mit.edu/2018/doctors-rely-gut-feelings-decision-making-0720.

This type of problem has caused accidents with self-driving cars, where the system either falsely identified a road hazard and braked unnecessarily and unexpectedly (false positive) or – in the case of one fatal accident in 2018 – it failed to identify a pedestrian crossing the road and did not brake at all (false negative).[60]

Vulnerability to spoofing

Yet another reason to be cautious about using computer vision algorithms for safety-critical tasks – especially in the absence of human verification – is their **vulnerability to tricking or spoofing by adversarial images or physical objects** (see also Section 4). Adding digital "noise" to an image that is not visible to the human eye can often cause a computer vision system to fail. More sophisticated adversarial techniques – for example changing a few pixels in a digital image – can trick an image recognition system into mistakes that a human would never make, and this has been demonstrated using adversarial physical objects in the real world. In a well-known example, researchers at the Massachusetts Institute of Technology tricked an image classification algorithm into classifying a 3-D printed turtle as a rifle, and a 3-D printed baseball as an espresso.[61]

Spoofing an algorithm with adversarial changes may not always be simple because these changes need to be robust enough to work when an image (or object) is rotated, zoomed, or filtered, and so an adversarial image that tricks one algorithm may not trick others. However, demonstrated adversarial tricking of image classification algorithms – whether in "white-box" attacks where the functioning of the AI algorithm is known or in "black-box" attacks where only the inputs and outputs of the machine learning system are known – **raises significant concerns about the reliability and predictability of these systems in real-world applications**. This is likely to be a particularly acute problem in the inherently adversarial environments of conflict, should such algorithms be used in weapon systems. Retaining a human on the loop for verification and authorization of a classification made by an algorithm – for example by checking against a live video feed – might provide a means of guarding against this problem to a certain extent (see Section 3.4), although researchers have recently shown that adversarial images may also fool humans.[62]

7. STANDARDS IN CIVILIAN AUTONOMOUS SYSTEMS

7.1 Safety-critical robotic systems

The development and use of autonomous systems for safety-critical tasks in the civilian sector raises the question as to whether there are lessons and standards for human control – and the human-machine relationship – that may be relevant to discussions of autonomy and AI weapon systems.

Because of the unpredictability problem, civilian autonomous robotic systems – and functions – generally perform only simple tasks in simple environments that present relatively low risks. However, some autonomous systems have been performing safety-critical tasks for some time, including industrial robots and aircraft autopilot systems. Others are in development and testing, such as self-driving cars. **There are similar questions about human control and supervision, procedures for emergency intervention and deactivation (including system fail-safes) and predictability and reliability**.

[60] A. Marshall, "False Positives: Self-Driving Cars and the Agony of Knowing What Matters", *Wired*, 29 May 2018: https://www.wired.com/story/self-driving-cars-uber-crash-false-positive-negative.
[61] A. Athalye *et al.*, *Synthesizing Robust Adversarial Examples*, Cornell University, v.3, 7 June 2018, https://arxiv.org/abs/1707.07397. M. Hutson, "A turtle – or a rifle? Hackers easily fool AIs into seeing the wrong thing", *Science*, 19 July 2018: http://www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.
[62] E. Ackerman, "Hacking the Brain With Adversarial Images", *IEEE Spectrum*, 28 February 2018, https://spectrum.ieee.org/the-human-os/robotics/artificial-intelligence/hacking-the-brain-with-adversarial-images.

Existing and emerging industry safety standards reflect these questions, although civilian standards often follow the development of technology rather than precede it, and are generally tailored to specific applications. Many existing safety standards for civilian robotics are hardware focused – largely because it is very difficult to verify software reliability – even though software is playing an increasingly important role in ensuring reliability and safety.

Industrial robots

Standards governing industrial robots are designed to limit the risk of accidents and injuries to workers in factories and warehouses.[63] For example, a US Robotic Industry Association standard has requirements for **emergency deactivation and speed control**; industrial robots must have a manually-activated emergency stop function that overrides all other controls, removes power from moving components, and remains active until reset manually. The standard also requires that loss of power to the robot's moving parts (e.g. a robot arm) must not lead to the release of a load that presents a hazard to operators. The standard also imposes limits on the movement of the robot and stipulates the use of **safeguarding measures – such as barriers or cages – that prevent human operators from entering an area where the robot could endanger them**. The robot must be designed to trigger an emergency stop if a person enters the safeguarded area while it is in autonomous mode.[64]

From the above we can see that in many situations where industrial robots are used, they are recognized as being dangerous by design, and measures are taken to reduce the risk of any contact with humans. While industrial robots are highly predictable as regards the repetitive tasks they perform, **unpredictability in consequences and associated risks arise from their interaction with humans**. However, the safety standard also contains measures to address the increasing use of "collaborative robots", which share a workspace with human operators. These include requirements for "monitored stop", where the robot stops when it detects an obstacle, speed reduction when a human operator is nearby and overall limits on the speed and force that the robot can generate. Other techniques to reduce risks to workers include ensuring that a moving robot always takes the same route.

International standards organizations are developing various standards for autonomous robotic systems. For example, the International Institute of Electrical and Electronics Engineers (IEEE) has an initiative on "Ethically Aligned Design" of autonomous and intelligent systems,[65] including development of the IEEE P7000 series of standards on: "Transparency of Autonomous Systems",[66] "Algorithmic Bias Considerations",[67] "Ontological Standard for Ethically Driven Robotics and Automation Systems"[68] and a "Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems".[69]

Aircraft

Stringent standards exist to ensure the safety of aircraft systems. The European Aviation Safety Agency and the US Federal Aviation Administration have similar standards for the reliability of aircraft components, including autopilot systems, requiring that they "**perform their intended functions under any foreseeable operating condition**",[70] and are designed so that:

[63] International Organization for Standardization, *ISO/TC 299 Robotics*, https://www.iso.org/committee/5915511/x/catalogue.
[64] ANSI/RIA, *Industrial Robots and Robot Systems – Safety Requirements*, ANSI/RIA R15.06-2012, 28 March 2013.
[65] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems: https://standards.ieee.org/industry-connections/ec/autonomous-systems.html.
[66] IEEE Standards Association, *P7001 – Transparency of Autonomous Systems*: https://standards.ieee.org/project/7001.html.
[67] IEEE Standards Association, *P7003 – Algorithmic Bias Considerations*: https://standards.ieee.org/project/7003.html.
[68] IEEE Standards Association, *P7007 – Ontological Standard for Ethically Driven Robotics and Automation Systems*: https://standards.ieee.org/project/7007.html.
[69] IEEE Standards Association, *P7009 – Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems*: https://standards.ieee.org/project/7009.html.
[70] US Department of Transportation, A*irworthiness standards: transport category airplanes, Equipment, systems, and installations, 14 CFR § 25.1309*, 1 January 2007: https://www.govinfo.gov/app/details/CFR-2007-title14-vol1/CFR-2007-title14-vol1-sec25-1309.

i) "any catastrophic failure condition is extremely improbable and does not result from a single failure"

ii) "any hazardous failure condition is extremely remote" and "any major failure condition is remote".

A "catastrophic failure" is one that would result in multiple failures, usually with the loss of the aircraft, and an "extremely improbable" failure condition is one that is so unlikely that it is not expected to occur during the entire operational life of all aircraft of one type.[71] Additionally, standards on **flight guidance systems – or autopilot – have specifications for human control and reliability**, specifically: "quick disengagement controls for the autopilot and autothrust functions must be provided for each pilot", "the autopilot must not create an unsafe condition when the flight crew applies an override force to the flight controls" and "under any condition of flight" the autopilot must not "produce unacceptable loads on the aeroplane" or "create hazardous deviations in the flight path".[72]

Road vehicles

Road vehicles are also subject to stringent safety standards – including those employing "automated driving systems" (the foundation of self-driving cars). These standards have been developed by standards bodies such as the International Organization for Standardization (ISO) and the Society of Automotive Engineers (SAE). The automotive industry aims for a zero per cent failure rate for electronic systems with an operating lifetime of up to 15 years and an ISO standard lays down Automotive Safety Integrity Levels for components based on a risk assessment that considers the severity of consequences, probability, and controllability (or ability of the user to avoid the harm).[73]

Standards covering human control, predictably and reliability for increasingly autonomous vehicles are still under development,[74] although the SAE has defined levels of automation to guide their development.[75] The US National Highway Traffic Safety Administration (NHTSA) has emphasized the need for "a robust design and validation process … with the goal of designing HAV [highly automated vehicle] systems free of unreasonable safety risks".[76] The NHTSA requires that the vehicle be able to alert the operator when it is not able to function, is malfunctioning, or the driver needs to take over. For systems "intended to operate without a human driver or occupant, the remote dispatcher or central control authority should be able to know the status of the HAV at all times".[77] The policy adds that "in cases of higher automation where a human driver may not be present, the HAV must be able to fall back into a minimal risk condition that may not include the driver" – fail-safe mode – which could include "automatically bringing the vehicle safely to a stop, preferably outside of an active lane of traffic".[78]

Although there are currently no genuinely self-driving cars in private use, they are being tested in a number of countries.[79] **Regulations generally stipulate that these vehicles can only be tested on public roads if there is a driver who can always take back control**. In California, for instance, regulations require that the driver is "either in immediate physical control of the vehicle or is actively monitoring the vehicle's operations and capable of taking over immediate physical control", and that the driver

[71] European Aviation Safety Agency, *Certification Specifications and Acceptable Means of Compliance for Large Aeroplanes, CS-25*, Amendment 12, 13 July 2012, CS 25.1309: https://www.easa.europa.eu/sites/default/files/dfu/CS-25%20Amendment%2014.pdf.
[72] *Ibid,* CS 25.1329.
[73] C. Hobbs, and P. Lee, "Understanding ISO 26262 ASILs", *Electronic Design*, 9 July 2013: https://www.electronicdesign.com/embedded/understanding-iso-26262-asils.
[74] International Organization for Standardization, *ISO/TC 204 Intelligent transport systems*: https://www.iso.org/committee/54706/x/catalogue.
[75] SAE, *Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles,* J3016_201806, 15 June 2018: https://www.sae.org/standards/content/j3016_201806.
[76] US Department of Transportation, *Federal Automated Vehicles Policy. Accelerating the Next Revolution in Roadway Safety*, NHTSA, September 2016, p. 20.
[77] *Ibid*, p. 22.
[78] *Ibid*, p. 30.
[79] A. Nunes, B. Reimer and J. Coughlin, "People must retain control of autonomous vehicles", *Nature*, 6 April 2018: https://www.nature.com/articles/d41586-018-04158-5.

"knows the limitations of the vehicle's autonomous technology and is capable of safely operating the vehicle in all conditions under which the vehicle is tested on public roads."[80]

Relevance of civilian standards to autonomous weapon systems

These safety standards for civilian applications may hold lessons for applications in armed conflict. While standards for human control of civilian systems are designed to ensure safety and avoid harm, standards for autonomous weapon systems must be designed to ensure they can be used with minimal risk of indiscriminate effects and other unintended consequences. In any case, **"safety" standards for human control in weapon systems should be at least as stringent as those for civilian applications**.

7.2 Governance of AI and machine learning

In parallel with the development of standards for physical autonomous systems, there is now increasing interest in the **underlying AI and machine-learning-based software that may both control physical robots and advise – or replace – humans in decisions** that are safety-critical, or present other significant consequences for human life and personal freedom. These have also brought ethical questions to the forefront of public debate, and a **common aspect of "AI principles"** developed and agreed by governments, scientists, ethicists, research institutes and technology companies **is the importance of the human element** in ensuring legal compliance and ethical acceptability.

7.2.1 AI principles

Future of Life Institute

The 2017 Asilomar AI Principles emphasize alignment with human values, compatibility with "human dignity, rights, freedoms and cultural diversity" and human control: "humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives".[81]

European Commission

The European Commission's High-Level Expert Group on Artificial Intelligence stressed the importance of "human agency and oversight", such that AI systems "support human autonomy and decision-making" and ensure human oversight through human-in-the-loop, human-on-the-loop or human-in-command approaches.[82]

OECD

The Organisation for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence – adopted in May 2019 by all 36 Member States with six other countries – highlight the importance of "human-centred values and fairness", specifying that users of AI "should implement mechanisms and safeguards, such as capacity for human determination, that are appropriate to the context and consistent with the state of art".[83]

[80] California, Department of Motor Vehicles, *Testing of Autonomous Vehicles with a Driver. Adopted Regulations for Testing of Autonomous Vehicles by Manufacturers*. Order to Adopt Title 13, Division 1, Chapter 1 Article 3.7 – Testing of Autonomous Vehicles, 26 February 2018.
[81] Future of Life Institute, *Asilomar AI Principles*, 2017: https://futureoflife.org/ai-principles.
[82] European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
[83] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449, 22 May 2019: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449. Adopted by the 36 Member States together with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania.

Beijing Academy of Artificial Intelligence

The Beijing AI Principles, adopted in May 2019 by a group of leading Chinese research institutes and technology companies, state that "continuous efforts should be made to improve the maturity, robustness, reliability, and controllability of AI systems" and encourage "explorations on human-AI coordination … that would give full play to human advantages and characteristics".[84]

Partnership on AI

The Partnership on AI – a multi-stakeholder initiative established by Apple, Amazon, DeepMind, Google, Facebook, IBM and Microsoft – highlighted best practice in safety-critical AI applications as an "urgent short-term question, with applications in medicine, transportation, engineering, computer security, and other domains".[85]

Individual companies

A number of individual technology companies have published AI principles highlighting the importance of human control,[86] especially for sensitive applications presenting the risk of harm,[87] and emphasizing that the "purpose of AI … is to augment – not replace – human intelligence".[88]

Google

Google has set out seven AI principles to guide its work, emphasizing social benefit, avoiding bias, ensuring safety, accountability and privacy. The principles require all their AI technologies to "be accountable to people" and "subject to appropriate human direction and control". The company has ruled out use in "applications that are likely to cause overall harm", "weapons or other technologies whose principal purpose or implementation is to cause or directly facilitate injury to people", "surveillance violating internationally accepted norms" and for applications "whose purpose contravenes widely accepted principles of international law and human rights."[89]

Google has called for guidance from governments and engagement from civil society on key concerns raised by AI, especially:

- explainability standards (or transparency)
- fairness (or bias)
- safety (or predictability and reliability)
- human-AI collaboration (or human control and supervision)
- liability.

On human-AI collaboration, the company affirms the necessity for a "human in the loop" in otherwise autonomous systems to address issues of safety and fairness (bias), and depending on the nature of the application. On the latter Google says **"it is likely there will always be sensitive contexts where society will want a human to make the final decision**, no matter how accurate an AI system is or the time/cost benefits of full automation."[90] The company has also highlighted the essential differences

[84] Beijing Academy of Artificial Intelligence (BAAI), *Beijing AI Principles*, 28 May 2019: https://baip.baai.ac.cn/en.
[85] The Partnership on AI*, op. cit.*
[86] Google, *AI at Google: Our principles*, 7 June 2018: https://www.blog.google/technology/ai/ai-principles. "We will design AI systems that provide appropriate opportunities for feedback, relevant explanations, and appeal. Our AI technologies will be subject to appropriate human direction and control."
[87] Microsoft, *Microsoft AI principles*, 2019: https://www.microsoft.com/en-us/ai/our-approach-to-ai. R. Sauer, *Six principles to guide Microsoft's facial recognition work*, Microsoft, 17 December 2018: https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work.
[88] IBM, *IBM's Principles for Trust and Transparency*, 30 May 2018: https://www.ibm.com/blogs/policy/trust-principles.
[89] Google, 2018, *op. cit.*
[90] Google, *Perspectives on Issues in AI Governance*, January 2019 p. 24: http://ai.google/perspectives-on-issues-in-AI-governance.

between humans and AI, stressing that "**machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it**."[91]

As regards regulation, Google said that "**governments may wish to identify red-line areas where human involvement is deemed imperative**" such as in "making legal judgments of criminality, or in making certain life-altering decisions about medical treatment" and asks for "broad guidance as to what human involvement should look like" in different contexts.[92] It concludes that some "contentious uses of AI" could represent such "a major and irrevocable shift in the scale of possible harm that could be inflicted" including "anything from a new kind of weapon to an application that fundamentally overhauls everyday norms (e.g. the ability to be anonymous in a crowd, or to trust in what you see)", that "additional rules would be of benefit".[93]

Microsoft

Microsoft has also been outspoken on sensitive applications of AI, in particular **facial recognition**, a technology which it is at the forefront of developing, calling for governments to adopt new regulation[94] and issuing principles to guide its work. Their **principle on accountability** says the company will encourage use of facial recognition technology "in a manner that ensures an **appropriate level of human control for uses that may affect people in consequential ways**", requiring a "human-in-the-loop" or "meaningful human review". Microsoft defines sensitive uses as those involving "**risk of bodily or emotional harm to an individual**, where an individual's employment prospects or ability to access financial services may be adversely affected, where there may be implications on human rights, or where an individual's personal freedom may be impinged."[95]

7.2.2 Relevance to discussions about the use of AI in armed conflict

Since applications of AI and machine learning in weapon systems – and in armed conflict more broadly – are likely to be among the most sensitive, **these broader governance discussions may be indicative of necessary constraints and of the type and degree of human control** and human-machine interaction that will be needed.

8. CONCLUSIONS

Autonomous weapon systems, which can select and attack targets without human intervention or self-initiate attacks raise concerns about loss of human control over the use of force. Like most States, the ICRC has called for human control to be retained to ensure compliance with international humanitarian law and ethical acceptability, and it has urged a focus on determining what human control means in practice.[96]

Based on the foregoing analysis, experience the civilian sector with autonomy, robotics and AI can yield insights for discussions about ensuring meaningful, effective and appropriate human control over weapon systems and the use of force, including in the following areas:

[91] *Ibid*, p. 21. "Such differences should be front of mind when thinking about the kind of tasks and settings in which to deploy an AI system to amplify and augment human capabilities."
[92] *Ibid*, p. 23.
93 *Ibid*, p. 29.
[94] B. Smith, *Facial recognition: It's time for action*, *Microsoft*, 6 December 2018, https://blogs.microsoft.com/on-the-issues/2018/12/06/facial-recognition-its-time-for-action.
[95] Microsoft, *Six principles for developing and deploying facial recognition technology*, December 2018: https://blogs.microsoft.com/wp-content/uploads/prod/sites/5/2018/12/MSFT-Principles-on-Facial-Recognition.pdf.
[96] ICRC, *ICRC Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems*, Geneva, 25–29 March 2019, *op. cit.* ICRC, *The Element of Human Control*, *op. cit.*

Human control

All autonomous (robotic) systems that operate without human intervention, based on interaction with their environment, **raise questions about how to ensure effective human control**.

Humans can exert some control over autonomous systems through human-on-the-loop supervision and intervention. This requires the operator to have situational awareness, enough time to intervene, and a mechanism through which to intervene (a communication link or physical controls) in order to take back control or deactivate the system. However, **human-on-the-loop control is not a panacea**, owing to such human-machine interaction problems as automation bias, lack of operator situational awareness and the moral buffer.

Predictability and reliability

It is **difficult to ensure and verify the predictability and reliability of an autonomous (robotic) system**. 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.

There is an important distinction between reliability – a measure of how often a system fails – and predictability – a measure of how the system will perform in a particular circumstance. 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. In the context of weapon systems, **both are important for ensuring compliance with international humanitarian law**.

AI and machine learning

AI algorithms – and especially machine learning systems – bring a new dimension of unpredictability to autonomous (robotic) systems. The "black box" manner in which most machine learning systems function today makes it difficult – and in most cases impossible – for the user to know how the system reaches its output. These systems 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 difficulty of testing and verifying the performance of autonomous systems.

Computer vision is an important application of machine learning, which is relevant to autonomous weapon systems. Most **computer vision systems use deep learning, of which the functioning is not predictable or transparent**, and which 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.

Standards for human control

We can learn lessons from **industry standards** for civilian safety-critical autonomous robotic systems, such as industrial robots, aircraft autopilot systems and self-driving cars, which are **stringent in their requirements** for human supervision, intervention, deactivation, predictability, reliability and operational constraints. **Leading civilian technology developers in AI and machine learning have also stressed the need to ensure human control** and judgement for sensitive uses – and to address safety and bias – especially where applications can have serious consequences for people's lives.

Towards limits on autonomy in weapon systems

These insights from the fields of autonomous systems, AI and robotics reinforce and expand 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 severe 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.

In establishing limits on autonomy in weapon systems it may be useful to consider sources of unpredictability that pose problems for human control and responsibility. **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 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 – 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 to control the critical functions of weapon systems, and it raises the questions of how to maintain human control and judgement in any use of machine learning in decision-support systems for targeting.[97]

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.

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.

[97] ICRC, *Artificial intelligence and machine learning in armed conflict: A human-centred approach*, *op. cit.*