International Review of the Red Cross (2020), **102** (913), 463–479.
Digital technologies and war
doi:10.1017/S1816383120000454

IRRC

REPORTS AND DOCUMENTS

Artificial intelligence and machine learning in armed conflict: A human-centred approach

Note: This is an edited version of a paper published by the ICRC in June 2019.

1. Introduction

At a time of increasing conflict and rapid technological change, the International Committee of the Red Cross (ICRC) needs both to understand the impact of new technologies on people affected by armed conflict and to design humanitarian solutions that address the needs of the most vulnerable.

The ICRC, like many organizations across different sectors and regions, is grappling with the implications of **artificial intelligence** (AI) and **machine learning** for its work. AI is the use of computer systems to carry out tasks – often associated with human intelligence – that require cognition, planning, reasoning or learning; and machine learning systems are AI systems that are "trained" on and "learn" from data, which ultimately define the way they function. Since these are software tools, or algorithms, that could be applied to many different tasks, the potential implications may be far-reaching and yet to be fully understood.

There are two broad – and distinct – areas of application of AI and machine learning in which the ICRC has a particular interest: first, its **use in the conduct of warfare** or in other situations of violence;[1] and second, its **use in humanitarian action** to assist and protect the victims of armed conflict.[2] This paper sets out the

ICRC's perspective on the use of AI and machine learning in armed conflict, the potential humanitarian consequences, and the associated legal obligations and ethical considerations that should govern its development and use. It also makes reference to the use of AI tools for humanitarian action, including by the ICRC.

2. The ICRC's approach to new technologies of warfare

The ICRC has a long tradition of assessing the implications of contemporary and near-future developments in armed conflict. This includes considering new means and methods of warfare; specifically, in terms of their compatibility with the rules of international humanitarian law (also known as the law of armed conflict, or the law of war) and the risks of adverse humanitarian consequences for protected persons.

The ICRC is not opposed to new technologies of warfare *per se*. Certain military technologies – such as those enabling greater precision in attacks – may assist conflict parties in minimizing the humanitarian consequences of war, in particular on civilians, and in ensuring respect for the rules of war. However, as with any new technology of warfare, precision technologies are not beneficial in themselves, and humanitarian consequences on the ground will depend on the way new weapons are used in practice. It is essential, therefore, to have a realistic assessment of new technologies that is informed by their technical characteristics *and* the way they are used, or are intended to be used.

Any new technology of warfare must be used, and must be capable of being used, in compliance with existing rules of international humanitarian law. This is a minimum requirement.[3] However, the unique characteristics of new technologies of warfare, the intended and expected circumstances of their use, and their foreseeable humanitarian consequences may raise questions of whether existing rules are sufficient or need to be clarified or supplemented, in light of the new technologies' foreseeable impact.[4] What is clear is that military applications of new and emerging technologies are not inevitable. They are choices made by

1 ICRC, "Expert Views on the Frontiers of Artificial Intelligence and Conflict", *ICRC Humanitarian Law and Policy Blog*, 19 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/19/expert-views-frontiers-artificial-intelligence-conflict.

2 ICRC, *Summary Document for UN Secretary-General's High-Level Panel on Digital Cooperation*, January 2019, available at: https://digitalcooperation.org/wp-content/uploads/2019/02/ICRC-Submission-UN-Panel-Digital-Cooperation.pdf.

3 States party to Additional Protocol I to the Geneva Conventions have an obligation to conduct legal reviews of new weapons during their development and acquisition, and prior to their use in armed conflict. For other States, legal reviews are a common-sense measure to help ensure that the State's armed forces can conduct hostilities in accordance with their international obligations.

4 ICRC, *International Humanitarian Law and the Challenges of Contemporary Armed Conflicts*, report for the 33rd International Conference of the Red Cross and Red Crescent, Geneva, October 2019 (ICRC Challenges Report 2019), pp. 18–29, available at: www.icrc.org/en/publication/4427-international-humanitarian-law-and-challenges-contemporary-armed-conflicts; 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 (ICRC Challenges Report 2015), pp. 38–47, available at: www.icrc.org/en/document/international-humanitarian-law-and-challenges-contemporary-armed-conflicts.

States which must be within the bounds of existing rules and must take into account potential humanitarian consequences for civilians and for combatants no longer taking part in hostilities, as well as broader considerations of "humanity" and "public conscience".[5]

3. Use of AI and machine learning by conflict parties

The ways in which parties to armed conflict – whether States or non-State armed groups – might use AI and machine learning in the conduct of warfare, and their potential implications, are not yet fully known. Nevertheless, there are at least **three overlapping areas that are relevant from a humanitarian perspective**, including for compliance with international humanitarian law.

3.1 Increasing autonomy in physical robotic systems, including weapons

One significant application is the use of digital **AI and machine learning tools to control physical military hardware**, in particular the increasing number of unmanned robotic systems – in the air, on land and at sea – with a wide range of sizes and functions. AI and machine learning may enable increasing autonomy in these robotic platforms, whether armed or unarmed, and whether controlling the whole system or specific functions such as flight, navigation, surveillance or targeting.

For the ICRC, **autonomous weapon systems** – weapon systems with autonomy in their "critical functions" of selecting and attacking targets – are an immediate concern from a humanitarian, legal and ethical perspective, given the risk of loss of human control over weapons and the use of force.[6] This loss of control raises risks for civilians, because of unpredictable consequences; legal questions,[7] because combatants must make context-specific judgements in carrying out attacks under international humanitarian law; and ethical concerns,[8] because human agency in decisions to use force is necessary to uphold moral

5 The "principles of humanity" and the "dictates of public conscience" are mentioned in Article 1(2) of Additional Protocol I and in the preamble of Additional Protocol II to the Geneva Conventions, referred to as the Martens Clause, which is part of customary international humanitarian law.

6 ICRC, Statements to the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 25–29 March 2019, available at: https://tinyurl.com/yyeadno3.

7 ICRC Challenges Report 2019, above note 4, pp. 29–31; Neil Davison, "Autonomous Weapon Systems under International Humanitarian Law", Perspectives on Lethal Autonomous Weapon Systems, United Nations Office for Disarmament Affairs Occasional Paper No. 30, November 2017, available at: www.icrc.org/en/document/autonomous-weapon-systems-under-international-humanitarian-law.

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

responsibility and human dignity. For these reasons, the ICRC has proposed practical elements of human control as the basis for internationally agreed limits on autonomy in weapon systems with a focus on the following:[9]

- **Controls on weapon parameters**, which can inform limits on types of autonomous weapon systems including the targets they are used against, as well as limits on their duration and geographical scope of operation, and requirements for deactivation and fail-safe mechanisms;
- **Controls on the environment**, which can inform limits on the situations and locations in which autonomous weapon systems may be used, notably in terms of the presence and density of civilians and civilian objects; and
- **Controls through human–machine interaction**, which can inform requirements for human supervision and ability to intervene and deactivate autonomous weapon systems, and requirements for predictable and transparent functioning.

It is important to recognize that **not all autonomous weapons incorporate AI and machine learning**; existing weapons with autonomy in their critical functions, such as air-defence systems with autonomous modes, generally use simple, rule-based control software to select and attack targets. However, **AI and machine learning software** – specifically of the type developed for "automatic target recognition" – **could form the basis of future autonomous weapon systems, bringing a new dimension of unpredictability to these weapons**, as well as concerns about lack of explainability and bias (see Section 5.2).[10] The same type of software might also be used in "decision support" applications for targeting, rather than directly to control a weapon system (see Section 3.3).

Conversely, not all military robotic systems using AI and machine learning are autonomous weapons, since the software might be used for control functions other than targeting, such as surveillance, navigation or flight. While, from the ICRC's perspective, autonomy in weapon systems – including AI-enabled systems – raises the most urgent questions, the use of AI and machine learning to increase autonomy in military hardware in general – such as in unmanned aircraft, land vehicles and sea vessels – may also raise questions of human–machine interaction and safety. Discussions in the civil sector about ensuring safety of autonomous vehicles – such as self-driving cars or drones – may hold lessons for their use in armed conflict (see also Section 3.3).

9 ICRC, *ICRC Commentary on the "Guiding Principles" of the CCW GGE on "Lethal Autonomous Weapons Systems"*, Geneva, July 2020, available at: https://documents.unoda.org/wp-content/uploads/2020/07/20200716-ICRC.pdf; Vincent Boulanin, Neil Davison, Netta Goussac and Moa Peldán Carlsson, *Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control*, ICRC and Stockholm International Peace Research Institute, June 2020, available at: www.icrc.org/en/document/limits-autonomous-weapons; ICRC, "The Element of Human Control", UN Doc. CCW/MSP/2018/WP.3, working paper, CCW Meeting of High Contracting Parties, 20 November 2018, available at: https://tinyurl.com/y3c96aa6.

10 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems under Agenda Item 6(b), Geneva, 27–31 August 2018, available at: https://tinyurl.com/y4cql4to.

3.2 New means of cyber and information warfare

The application of **AI and machine learning to the development of cyber weapons or capabilities** is another important area. Not all cyber capabilities incorporate AI and machine learning. However, these technologies are expected to **change the nature of both capabilities to defend against cyber attacks and capabilities to attack**. For example, AI and machine learning-enabled cyber capabilities could automatically search for vulnerabilities to exploit, or defend against cyber attacks while simultaneously automatically launching counter-attacks. These types of developments could increase the scale, and change the nature and perhaps the severity, of attacks.[11] Some of these systems might even be described as "digital autonomous weapons", potentially raising questions about human control similar to those that apply to physical autonomous weapons.[12]

The ICRC's focus with respect to cyber warfare remains on ensuring that existing international humanitarian law rules are upheld in any cyber attacks in armed conflict, and that the particular challenges in ensuring the protection of civilian infrastructure and services are addressed by those carrying out or defending against such attacks,[13] in order to minimize the human cost.[14]

A related application of AI and machine learning in the digital sphere is the **use of these tools for information warfare**, in particular the creation and spreading of false information with intent to deceive – i.e., **disinformation** – as well as the spreading of false information without such intent – i.e., **misinformation**. Not all involve AI and machine learning, but these technologies seem set to change the nature and scale of the manipulation of information in warfare as well as the potential consequences. AI-enabled systems have been widely used to produce fake information – whether text, audio, photos or video – which is increasingly difficult to distinguish from real information. Use of these systems by conflict parties to amplify age-old methods of propaganda in order to manipulate opinion and influence decisions could have significant implications on the ground.[15] For the ICRC, there are concerns that civilians might, as a result of digital disinformation or misinformation, be subject to arrest or ill-treatment, discrimination or denial of access to essential services, or attacks on their person or property.[16]

11 Miles Brundage *et al.*, *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*, Future of Humanity Institute, Oxford, February 2018.
12 United Nations Institute for Disarmament Research (UNIDIR), *The Weaponization of Increasingly Autonomous Technologies: Autonomous Weapon Systems and Cyber Operations*, 2017.
13 By asserting that international humanitarian law applies to cyber operations, the ICRC is in no way condoning cyber warfare, nor is it condoning the militarization of cyberspace: ICRC Challenges Report 2015, above note 4, pp. 38–44.
14 ICRC, *The Potential Human Cost of Cyber Operations*, report of an expert meeting, Geneva, May 2019, available at: www.icrc.org/en/document/potential-human-cost-cyber-operations.
15 Steven Hill and Nadia Marsan, "Artificial Intelligence and Accountability: A Multinational Legal Perspective", in *Big Data and Artificial Intelligence for Military Decision Making*, STO Meeting Proceedings STO-MP-IST-160, NATO, 2018.
16 ICRC, *Symposium Report: Digital Risks in Situations of Armed Conflict*, March 2019, p. 9, available at: www.icrc.org/en/event/digital-risks-symposium.

3.3 Changing nature of decision-making in armed conflict

Perhaps the broadest and most far-reaching application is the use of **AI and machine learning for decision-making**, enabling widespread collection and analysis of data sources in order to identify people or objects, assess patterns of life or behaviour, make recommendations for military strategy or operations, or make predictions about future actions or situations.

These **"decision support"** or **"automated decision-making" systems are effectively an expansion of intelligence, surveillance and reconnaissance tools**, using AI and machine learning to automate the analysis of large data sets in order to provide "advice" to humans in making particular decisions, or to automate both the analysis and the subsequent initiation of a decision or action by the system. Relevant AI and machine learning applications include pattern recognition, natural language processing, image recognition, facial recognition and behaviour recognition. The **possible use of these systems is extremely broad**,[17] from decisions about who – or what – to attack and when,[18] to decisions about who to detain and for how long,[19] to decisions about military strategy – even on use of nuclear weapons[20] – and specific operations, including attempts to predict or pre-empt adversaries.[21] Depending on their use or misuse – and the capabilities and limitations of the technology – these decision-making applications could lead to increased risks for civilian populations.

AI and machine learning-based **decision support systems** may enable better decisions by humans in conducting hostilities in compliance with international humanitarian law and minimizing risks for civilians by facilitating quicker and more widespread collection and analysis of available information. However, over-reliance on the same algorithmically generated analyses, or predictions, might also facilitate worse decisions or violations of international humanitarian law and exacerbate risks for civilians, especially given the current limitations of the technology, such as unpredictability, lack of explainability and bias (see Section 5.2).

From a humanitarian perspective, a **very wide range of different AI-mediated – or AI-influenced – decisions by conflict parties could be relevant**, especially where they pose risks of injury or death to persons or destruction of objects, and where the decisions are governed by specific rules of international humanitarian law. For example, the use of AI and machine learning for **targeting decisions in armed conflict**, where there are serious consequences for life, will require specific considerations to ensure humans remain in a position to make

17 Dustin A. Lewis, Gabriella Blum and Naz K. Modirzadeh, *War-Algorithm Accountability*, Harvard Law School Program on International Law and Armed Conflict, August 2016.

18 United States, "Implementing International Humanitarian Law in the Use of Autonomy in Weapon Systems", working paper, CCW Group of Governmental Experts, March 2019.

19 Ashley Deeks, "Predicting Enemies", Virginia Public Law and Legal Theory Research Paper No. 2018-21, March 2018, available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3152385.

20 Vincent Boulanin (ed.), *The Impact of Artificial Intelligence on Strategic Stability and Nuclear Risk*, Vol. 1: *Euro-Atlantic Perspectives*, Stockholm International Peace Research Institute, Stockholm, May 2019.

21 S. Hill and N. Marsan, above note 15.

the context-based judgements required for compliance with the legal rules on the conduct of hostilities (see Section 5). An AI system used to directly initiate an attack (rather than producing an analysis, or "advice", for human decision-makers) would effectively be considered an autonomous weapon system, raising similar issues (see Section 3.1).

The use of decision support and automated decision-making systems may also raise **legal and ethical questions for other applications, such as decisions on detention in armed conflict**, which also have serious consequences for people's lives and are governed by specific rules of international humanitarian law. Here there are parallels with discussions in the civil sector about the role of human judgement, and issues of bias and inaccuracy, in risk-assessment algorithms used by the police in decisions on arrest, and in the criminal justice system for decisions on sentencing and bail.[22]

More broadly, these types of AI and machine learning tools might lead to an increasing **personalization of warfare** (with parallels to the personalization of services in the civilian world), with digital systems bringing together personally identifiable information from multiple sources – including sensors, communications, databases, social media and biometric data – to form an algorithmically generated determination about a person, their status and their targetability, or to predict their future actions.

In general, potential humanitarian consequences – **digital risks** – for civilian populations from misuse of AI-enabled **digital surveillance, monitoring and intrusion** technologies could include being targeted, arrested, facing ill-treatment, having their identity stolen and being denied access to services, having assets stolen or suffering from psychological effects from the fear of being under surveillance.[23]

4. Use of AI and machine learning for humanitarian action

The ways in which AI and machine learning might be used for humanitarian action, including by the ICRC, are also likely to be very broad. These tools are being explored by humanitarian organizations for environment scanning, monitoring and analysis of public sources of data in specific operational contexts – applications that could help **inform assessments of humanitarian needs**, such as the type of assistance needed (food, water, shelter, economic, health) and where it is needed.

Similar AI-enabled data aggregation and analysis tools might be used to help **understand humanitarian consequences** on the ground, including civilian protection needs – for example, tools for image, video or other pattern analysis to assess damage

22 Lorna McGregor, "The Need for Clear Governance Frameworks on Predictive Algorithms in Military Settings", *ICRC Humanitarian Law and Policy Blog*, 28 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/28/need-clear-governance-frameworks-predictive-algorithms-military-settings; AI Now Institute, *AI Now Report 2018*, New York University, December 2018, pp. 18–22.

23 ICRC, above note 16, p. 8.

to civilian infrastructure, patterns of population displacement, viability of food crops, or the degree of weapon contamination (unexploded ordnance). These systems might also be used to analyze images and videos in order to detect and assess the conduct of hostilities, and the resulting humanitarian consequences.

The ICRC, for example, has developed **environment scanning dashboards** using AI and machine learning to capture and analyze large volumes of data to inform and support its humanitarian work in specific operational contexts, including using predictive analytics to help determine humanitarian needs.

A wide range of humanitarian services might benefit from the application of AI and machine learning tools for specific tasks. For example, there is interest in technologies that could **improve identification of missing persons**, such as AI-based facial recognition and natural language processing for name matching; the ICRC has been exploring the use of these technologies to support the work of its Central Tracing Agency in reuniting family members separated by conflict. It is also exploring the use of AI and machine learning-based **image analysis and pattern recognition for satellite imagery**, whether to map population density in support of infrastructure assistance projects in urban areas or to complement its documentation of respect for international humanitarian law as part of its civilian protection work.

These **applications for humanitarian action also bring potential risks**, as well as legal and ethical questions, in particular with respect to data protection, privacy, human rights, accountability and ensuring human involvement in decisions with significant consequences for people's lives and livelihoods. Any applications for humanitarian action must be designed and used under the principle of **"do no harm"** in the digital environment, and respect the right to privacy, including as it relates to personal data protection.

The ICRC will also ensure that the **core principles and values of neutral, independent and impartial humanitarian action** are reflected in the design and use of AI and machine learning applications it employs, taking into account a realistic assessment of the capabilities and limitations of the technology (see Section 5.2). The ICRC has led – with the Brussels Privacy Hub – an initiative on data protection in humanitarian action to develop guidance on the use of new technologies, including AI and machine learning, in the humanitarian sector in a way that maximizes the benefits without losing sight of these core considerations. The second edition of the ICRC/Brussels Privacy Hub *Handbook on Data Protection in Humanitarian Action* was published in May 2020.[24]

5. A human-centred approach

As a humanitarian organization working to protect and assist people affected by armed conflict and other situations of violence, deriving its mandate from international humanitarian law and guided by the Fundamental Principle of

24 ICRC and Brussels Privacy Hub, *Handbook on Data Protection in Humanitarian Action*, 2nd ed., Geneva, May 2020, available at: www.icrc.org/en/data-protection-humanitarian-action-handbook.

humanity,[25] the **ICRC believes it is critical to ensure a genuinely human-centred approach to the development and use of AI and machine learning**. This starts with consideration of the obligations and responsibilities of humans and what is required to ensure that the use of these technologies is compatible with international law, as well as societal and ethical values.

5.1 Ensuring human control and judgement

The ICRC believes it is **essential to preserve human control over tasks and human judgement in decisions that may have serious consequences** for people's lives in armed conflict, especially where these tasks and decisions pose risks to life, and where they are governed by specific rules of international humanitarian law. **AI and machine learning systems must be used to serve human actors, and as tools to augment human decision-makers, not replace them**. Given that these technologies are being developed to perform tasks that would ordinarily be carried out by humans, there is an inherent tension between the pursuit of AI and machine learning applications and the centrality of the human being in armed conflict, which will need continued attention.

Human control and judgement will be particularly important for tasks and decisions that can lead to injury or loss of life, or damage to, or destruction of, civilian infrastructure. These will likely raise the most serious legal and ethical questions, and may demand policy responses, such as new rules and regulations. **Most significant are decisions on the use of force, determining who and what is targeted and attacked in armed conflict**. However, a much wider range of tasks and decisions to which AI might be applied could also have serious consequences for those affected by armed conflict, such as decisions on arrest and detention. In considering the use of AI for sensitive tasks and decisions, there may be lessons from broader discussions in the civil sector about the governance of "safety-critical" AI applications – those whose failure can lead to injury or loss of life, or serious damage to property or the environment.[26]

Another area of tension is the **discrepancy between humans and machines in the speed at which they carry out different tasks**. Since humans are the legal – and moral – agents in armed conflict, the technologies and tools they use to conduct warfare must be designed and used in a way that enables combatants to fulfil their legal and ethical obligations and responsibilities. This may have significant implications for AI and machine learning systems that are used in decision-

25 ICRC and International Federation of Red Cross and Red Crescent Societies, *The Fundamental Principles of the International Red Cross and Red Crescent Movement: Ethics and Tools for Humanitarian Action*, Geneva, November 2015, available at: https://shop.icrc.org/les-principes-fondamentaux-de-la-croix-rouge-et-du-croissant-rouge-2757.html.

26 See, for example, the Partnership on AI's focus on the safety of AI and machine learning technologies as "an urgent short-term question, with applications in medicine, transportation, engineering, computer security, and other domains hinging on the ability to make AI systems behave safely despite uncertain, unanticipated, and potentially adversarial environments". Partnership on AI, "Safety-Critical AI: Charter", 2018, available at: www.partnershiponai.org/working-group-charters-guiding-our-exploration-of-ais-hard-questions.

making; in order to preserve human judgement, systems may need to be designed and used to inform decision-making at "human speed", rather than accelerating decisions to "machine speed" and beyond human intervention.

Legal basis for human control in armed conflict

For conflict parties, **human control over AI and machine learning applications employed as means and methods of warfare is required to ensure compliance with the law**. The rules of international humanitarian law are addressed to humans. It is humans that comply with and implement the law, and it is humans who will be held accountable for violations. In particular, combatants have a unique obligation to make the judgements required of them by the international humanitarian law rules governing the conduct of hostilities, and this responsibility cannot be transferred to a machine, a piece of software or an algorithm.

These rules require context-specific judgements to be taken by those who plan, decide upon and carry out attacks, in order to: ensure **distinction** – between military objectives, which may lawfully be attacked, and civilians or civilian objects, which must not be attacked; ensure **proportionality** – in terms of ensuring that the incidental civilian harm expected from an attack will not be excessive in relation to the concrete and direct military advantage anticipated; and enable **precautions in attack** – so that risks to civilians can be further minimized.

Where AI systems are used in attacks – whether as part of physical or cyber-weapon systems, or in decision support systems – **their design and use must enable combatants to make these judgements**.[27] With respect to autonomous weapon systems, the States party to the Convention on Certain Conventional Weapons (CCW), have recognized that "human responsibility" for the use of weapon systems and the use of force "must be retained",[28] and many States, international organizations – including the ICRC – and civil society organizations have stressed the requirement for human control to ensure compliance with international humanitarian law and compatibility with ethical values.[29]

Beyond the use of force and targeting, the potential use of AI systems for other decisions governed by specific rules of international humanitarian law will likely require careful consideration of necessary human control, and judgement, such as in detention.[30]

27 ICRC, above note 6.

28 United Nations, *Report of the 2018 session of the Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems*, UN Doc. CCW/GGE.1/2018/3, 23 October 2018, Sections III.A.26(b), III.C.28(f), available at: http://undocs.org/en/CCW/GGE.1/2018/3.

29 See, for example, the statements delivered at the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 25–29 March 2019, available at: https://tinyurl.com/yyeadno3.

30 Tess Bridgeman, "The Viability of Data-Reliant Predictive Systems in Armed Conflict Detention", *ICRC Humanitarian Law and Policy Blog*, 8 April 2019, available at: https://blogs.icrc.org/law-and-policy/2019/04/08/viability-data-reliant-predictive-systems-armed-conflict-detention.

Ethical basis for human control

Emerging applications of AI and machine learning have also brought ethical questions to the forefront of public debate. **A common aspect of general "AI principles"** developed and agreed by governments, scientists, ethicists, research institutes and technology companies **is the importance of the human element** to ensure legal compliance and ethical acceptability.

For example, 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".[31] The European Commission's High-Level Expert Group on Artificial Intelligence stressed the importance of "human agency and oversight", such that AI systems should "support human autonomy and decision-making", and of ensuring human oversight through human-in-the-loop, human-on-the-loop or human-in-command approaches.[32] The Organisation for Economic Co-operation and Development (OECD) Principles on Artificial Intelligence – adopted in May 2019 by all thirty-six member States, together with Argentina, Brazil, Colombia, Costa Rica, Peru and Romania – 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".[33] 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".[34] A number of individual technology companies have also published AI principles highlighting the importance of human control,[35] especially for sensitive applications presenting the risk of harm,[36] and emphasizing that the "purpose of AI … is to augment – not replace – human intelligence".[37]

31 Future of Life Institute, "Asilomar AI Principles", 2017, available at: https://futureoflife.org/ai-principles.
32 European Commission, *Ethics Guidelines for Trustworthy AI*, High-Level Expert Group on Artificial Intelligence, 8 April 2019, pp. 15–16, available at: https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.
33 OECD, "Recommendation of the Council on Artificial Intelligence", OECD/LEGAL/0449, 22 May 2019, available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.
34 Beijing Academy of Artificial Intelligence, "Beijing AI Principles", 28 May 2019, available at: https://baip.baai.ac.cn/en.
35 Google, "AI at Google: Our Principles", *The Keyword*, 7 June 2018, available at: 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."
36 Microsoft, "Microsoft AI Principles", 2019, available at: www.microsoft.com/en-us/ai/our-approach-to-ai; Rich Sauer, "Six Principles to Guide Microsoft's Facial Recognition Work", *Microsoft Blog*, 17 December 2018, available at: https://blogs.microsoft.com/on-the-issues/2018/12/17/six-principles-to-guide-microsofts-facial-recognition-work.
37 IBM, "IBM's Principles for Trust and Transparency", *THINKPolicy Blog*, 30 May 2018 available at: www.ibm.com/blogs/policy/trust-principles.

Some governments are also developing AI principles for the military. For example, the US Department of Defense (DoD), which called for the "human-centered" adoption of AI in its 2018 AI Strategy,[38] tasked its Defense Innovation Board with providing recommendations. Foremost among them was that "[h]uman beings should exercise appropriate levels of judgment and remain responsible" for any use of AI.[39] This informed the first of five DoD principles adopted in early 2020, which states that AI must be "[r]esponsible. DoD personnel will exercise appropriate levels of judgment and care, while remaining responsible for the development, deployment, and use of AI capabilities".[40] In France, the Ministry of Defence has committed to the use of AI in line with three guiding principles – compliance with international law, maintaining sufficient human control, and ensuring permanent command responsibility – and has established a Ministerial Ethics Committee to address emerging technologies.[41]

In the ICRC's view, preserving **human control** over tasks and **human judgement** in decisions that have serious consequences for people's lives will also be **essential to preserve a measure of humanity in warfare**. **The ICRC has stressed the need to retain human agency over decisions to use force in armed conflict**,[42] a view which derives from broader ethical considerations of humanity, moral responsibility, human dignity and the dictates of public conscience.[43]

However, ethical considerations of human agency may have broader applicability to other uses of AI and machine learning in armed conflict and other situations of violence. There are perhaps **lessons from wider societal discussions about sensitive applications of dual-use AI and machine learning technologies**, especially for safety-critical applications, and associated proposals for governance by scientists and developers in the private sector. Google, for example, has said that there may be "sensitive contexts where society will want a human to make the final decision, no matter how accurate an AI system", and that fully delegating high-stakes decisions to machines – such as legal judgements of criminality or life-altering decisions about medical treatment – "may fairly be seen as an affront to human dignity".[44] Microsoft, in considering AI-based facial recognition, has emphasized ensuring "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" for sensitive uses such as those involving

38 DoD, *Summary of the 2018 Department of Defense Artificial Intelligence Strategy*, 2019.

39 DoD, Defense Innovation Board, *AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense*, 31 October 2019.

40 DoD, "DOD Adopts Ethical Principles for Artificial Intelligence", news release, 24 February 2020, available at: www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/.

41 French Ministry of Defence, "Florence Parly Wants High-Performance, Robust and Properly Controlled Artificial Intelligence", *Actualités*, 10 April 2019, available at: www.defense.gouv.fr/english/actualites/articles/florence-parly-souhaite-une-intelligence-artificielle-performante-robuste-et-maitrisee.

42 ICRC, *ICRC Strategy 2019–2022*, Geneva, 2018, p. 15, available at: www.icrc.org/en/publication/4354-icrc-strategy-2019-2022.

43 ICRC, above note 8, p. 22.

44 Google, *Perspectives on Issues in AI Governance*, January 2019, pp. 23–24, available at: http://ai.google/perspectives-on-issues-in-AI-governance.

"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".[45] Since applications in armed conflict are likely to be among the most sensitive, these broader discussions may hold insights for necessary constraints on AI applications.

Preserving **human control and judgement will be an essential component** for ensuring legal compliance and mitigating ethical concerns raised by certain applications of AI and machine learning. **But it will not, in itself, be sufficient to guard against potential risks** without proper consideration of human–machine interaction issues such as: **situational awareness** (knowledge of the state of the system at the time of human intervention); **time available** for effective human intervention; **automation bias** (risk of human overtrust in the system); and the **moral buffer** (risk of humans transferring responsibility to the system).[46] Further, ensuring meaningful and effective human control and judgement will require careful consideration of both the capabilities and the limitations of AI and machine learning technologies.

5.2 Understanding the technical limitations of AI and machine learning

While much is made of the new capabilities offered by AI and machine learning, a **realistic assessment of the capabilities and limitations of these technologies is needed**, especially if they are to be used for applications in armed conflict. This should start with an acknowledgement that in using AI and machine learning for certain tasks or decisions, we are not replacing like with like. It requires an **understanding of the fundamental differences in the way humans and machines do things, as well as their different strengths and weaknesses**; humans and machines do things differently, and they do different things. We must be clear that, as inanimate objects and tools for use by humans, "machines will never be able to bring a genuine humanity to their interactions, no matter how good they get at faking it".[47]

With this in mind, there are several technical issues that demand caution in considering applications in armed conflict (and indeed for humanitarian action). **AI, and especially machine learning, brings concerns about unpredictability and unreliability** (or safety),[48] **lack of transparency** (or explainability), **and bias**.[49]

45 R. Sauer, above note 36: "We will encourage and help our customers to deploy facial recognition technology in a manner that ensures an appropriate level of human control for uses that may affect people in consequential ways."
46 ICRC, above note 8, p. 13.
47 Google, above note 44, p. 22.
48 Dario Amodei *et al.*, *Concrete Problems in AI Safety*, Cornell University, Ithaca, NY, 2016, available at: https://arxiv.org/abs/1606.06565.
49 ICRC, *Autonomy, Artificial Intelligence and Robotics: Technical Aspects of Human Control*, report of an expert meeting, Geneva, August 2019, available at: www.icrc.org/en/document/autonomy-artificial-intelligence-and-robotics-technical-aspects-human-control.

Rather than following a pre-programmed sequence of instructions, **machine learning systems build their own rules based on the data they are exposed to** – whether training data or through trial-and-error interaction with their environment. **As a result, they are much more unpredictable** than pre-programmed systems in terms of how they will function (reach their output) in a given situation (with specific inputs), and their functioning is highly dependent on the quantity and quality of available data for a specific task. For the developer it is difficult to know when the training is complete, or even what the system has learned. The same machine learning system may respond differently even when exposed to the same situation, and some systems may lead to unforeseen solutions to a particular task.[50] These core problems are exacerbated when the system continues to "learn" and change its model after deployment for a specific task. The unpredictable nature of machine learning systems, which can be an advantage in solving tasks, may not be a problem for benign tasks, such as playing a board game,[51] but it may be a significant concern for applications in armed conflict, such as autonomous weapon systems, cyber warfare and decision support systems (see Sections 3.1–3.3).

Complicating matters further, many machine learning systems are **not transparent**; **they produce outputs that are not explainable**. This "black box" nature makes it difficult – and, in many cases, currently impossible – for the user to understand *how* and *why* the system reaches its output from a given input; in other words, there is a lack of explainability and interpretability.

These issues of unpredictability and lack of explainability make **establishing trust in AI and machine learning systems a significant challenge**. An additional problem for trust is **bias**, which can have many facets, whether reinforcing existing human biases or introducing new ones in the design and/or use of the system. A common form is bias from training data, where limits in the quantity, quality and nature of available data to train an algorithm for a specific task can introduce bias into the functioning of the system relative to its task. This will likely be a significant issue for applications in armed conflict, where high-quality, representative data for specific tasks is scarce. However, other forms of bias can derive from the weighting given to different elements of data by the system, or to its interaction with the environment during a task.[52]

Concerns about unpredictability, lack of transparency or explainability, and bias have been documented in various applications of AI and machine learning, for example in image recognition,[53] facial recognition[54] and automated

50 Joel Lehman *et al.*, *The Surprising Creativity of Digital Evolution: A Collection of Anecdotes from the Evolutionary Computation and Artificial Life Research Communities*, Cornell University, Ithaca, NY, 2018, available at: https://arxiv.org/abs/1803.03453.

51 David Silver *et al.*, "Mastering the Game of Go without Human Knowledge", *Nature*, Vol. 550, No. 7676, 19 October 2017.

52 UNIDIR, *Algorithmic Bias and the Weaponization of Increasingly Autonomous Technologies: A Primer*, 2018.

53 Matthew Hutson, "A Turtle – or a Rifle? Hackers Easily Fool AIs into Seeing the Wrong Thing", *Science*, 19 July 2018, available at: www.sciencemag.org/news/2018/07/turtle-or-rifle-hackers-easily-fool-ais-seeing-wrong-thing.

54 AI Now Institute, above note 22, pp. 15–17.

decision-making systems.[55] Another fundamental issue with applications of AI and machine learning, such as computer vision, is **the semantic gap**, which shows that humans and machines carry out tasks very differently.[56] A computer-vision algorithm trained on images of particular subjects may be able to identify and classify those subjects in a new image. However, the algorithm has no understanding of the *meaning* or *concept* of that subject, which means it can make mistakes that a human never would, such as classifying an object as something completely different and unrelated. This would obviously raise serious concerns in certain applications in armed conflict, such as in autonomous weapon systems or decision support systems for targeting (see Sections 3.1 and 3.3).

The use of AI and machine learning in armed conflict will likely be even more difficult to trust in situations where it can be assumed that adversaries will apply countermeasures such as trying to trick or spoof each other's systems. **Machine learning systems are particularly vulnerable to adversarial conditions**, whether modifications to the environment designed to fool the system or the use of another machine learning system to produce adversarial images or conditions (a generative adversarial network, or GAN). In a well-known example, researchers tricked an image classification algorithm into identifying a 3D-printed turtle as a "rifle", and a 3D-printed baseball as an "espresso".[57] The risks of this type of problem are also clear should an AI-based image recognition system be used in weapon systems or for targeting decisions.

6. Conclusions and recommendations

AI and machine learning systems could have **profound implications for the role of humans in armed conflict**, especially in relation to: increasing autonomy of weapon systems and other unmanned systems; new forms of cyber and information warfare; and, more broadly, the nature of decision-making. In the view of the ICRC, governments, militaries and other relevant actors in armed conflict must pursue a genuinely **human-centred approach to the use of AI and machine learning systems based on legal obligations and ethical responsibilities**. The use of AI in weapon systems must be approached with great caution.

As a general principle, it is **essential to preserve human control and judgement in applications of AI and machine learning for tasks and in decisions that may have serious consequences for people's lives**, especially where these tasks and decisions pose risks to life, and where they are governed by specific rules of international humanitarian law. **AI and machine learning systems remain tools that must be used to serve human actors, and augment human decision-makers, not replace them**.

55 *Ibid.*, pp. 18–22.
56 Arnold W. M. 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, 2000.
57 M. Hutson, above note 53.

Ensuring human control and judgement in AI-enabled physical and digital systems that present such risks will be **needed for compliance with international humanitarian law and, from an ethical perspective, to preserve a measure of humanity in armed conflict**. In order for humans to meaningfully play their role, these systems may need to be designed and used to **inform decision-making at human speed, rather than accelerating decisions to machine speed** and beyond human intervention. These considerations may ultimately lead to constraints in the design and use of AI and machine learning systems to allow for meaningful and effective human control and judgement, based on legal obligations and ethical responsibilities.

An overall principle of human control and judgement is an essential component, but it is not sufficient in itself to guard against the potential risks of AI and machine learning in armed conflict. **Other related aspects to consider** will be ensuring: **predictability** and **reliability** – or safety – in the operation of the system and the consequences that result; **transparency** – or **explainability** – in how the system functions and why it reaches a particular output; and **lack of bias** – or fairness – in the design and use of the system. These issues will need to be addressed in order to **build trust** in the use of a given system, including through **rigorous testing in realistic environments** before being put into operation.[58]

The nature of human–AI interaction required will likely depend on ethical considerations and the particular rules of international humanitarian law and other applicable law that apply in the circumstances. Therefore, **general principles may need to be supplemented by specific principles, guidelines or rules on the use of AI and machine learning for specific applications and in particular circumstances**.

In the ICRC's view, one of the most pressing concerns is the relationship between humans and machines in decisions to kill, injure, damage or destroy, and the **critical importance of ensuring human control over weapon systems and the use of force** in armed conflict. With increasingly autonomous weapon systems, whether AI-enabled or not, there is a risk of effectively leaving these decisions to sensors and algorithms, a prospect that raises legal and ethical concerns which must be addressed with some urgency.

The ICRC has proposed key elements of human control necessary to comply with international humanitarian law and satisfy ethical concerns as a basis for internationally agreed limits on autonomy in weapon systems, including controls on weapon parameters, controls on the environment and

58 Netta Goussac, "Safety Net or Tangled Web: Legal Reviews of AI in Weapons and War-fighting", *ICRC Humanitarian Law and Policy Blog*, 18 April 2019, available at: https://blogs.icrc.org/law-and-policy/2019/04/18/safety-net-tangled-web-legal-reviews-ai-weapons-war-fighting; Dustin A. Lewis, "Legal Reviews of Weapons, Means and Methods of Warfare Involving Artificial Intelligence: 16 Elements to Consider", *ICRC Humanitarian Law and Policy Blog*, 21 March 2019, available at: https://blogs.icrc.org/law-and-policy/2019/03/21/legal-reviews-weapons-means-methods-warfare-artificial-intelligence-16-elements-consider.

controls through human–machine interaction.[59] It is clear to the ICRC that limits are needed on the types of autonomous weapons used and the situations in which they are used.[60]

This **human control-based approach** to autonomous weapon systems **would also be pertinent to broader applications of AI and machine learning in decision-making in armed conflict**, in particular where there are significant risks for human life and specific rules of international humanitarian law that apply, such as the use of decision support systems for targeting and detention.

59 ICRC, *Commentary on the "Guiding Principles"*, above note 9; ICRC, "The Element of Human Control", above note 9; V. Boulanin *et al.*, above note 9.

60 ICRC, Statement to the CCW Group of Governmental Experts on Lethal Autonomous Weapons Systems, Geneva, 21–25 September 2020, available at: https://documents.unoda.org/wp-content/uploads/2020/09/20200921-ICRC-General-statement-CCW-GGE-LAWS-Sep-2020.pdf.