SUBMISSION TO THE UNITED NATIONS SECRETARY-GENERAL ON ARTIFICIAL INTELLIGENCE IN THE MILITARY DOMAIN

RE: ODA/2025-00029/AIMD

The International Committee of the Red Cross (ICRC) welcomes the opportunity to submit its views for consideration by the United Nations secretary-general, in accordance with resolution 79/239 on "Artificial intelligence in the military domain and its implications for international peace and security", adopted by the General Assembly on 24 December 2024, which requested the secretary-general to seek views on "the opportunities and challenges posed to international peace and security by the application of artificial intelligence in the military domain".

The ICRC is a neutral, impartial and independent humanitarian organization. Through the Geneva Conventions of 1949 (Geneva Conventions) and other international legal instruments, the ICRC is mandated by States to protect and assist people affected by armed conflict. The ICRC also endeavours to prevent suffering by promoting and strengthening (including, where necessary, through contributing to the development of) international humanitarian law (IHL) and universal humanitarian principles.

The ICRC makes this submission based on its 160 years' experience of humanitarian action, during which time its staff have witnessed the significant humanitarian consequences of armed conflict, whether that be direct or indirect harm to people, objects, communities and societies.

Over the course of its history, the ICRC has played a significant role in the development of many of the IHL rules regulating the use of means and methods of warfare. The ICRC's work related to the regulation of weapons, means and methods of warfare is always driven by an "effects-based" approach. This means that we assess the actual, or – in the case of new weapons, means or methods not yet used – the foreseeable effects of their use, both on civilians and combatants. We then raise our concerns regarding particular weapons, means and methods of warfare that pose legal or ethical challenges, or present other risks of harm to those affected by armed conflict.

The recommendations that the ICRC makes in this submission are in line with its long-standing mandate and practice of promoting respect for and the development of IHL, including its application to new technologies of warfare. In line with Action 27(d) of the Pact for the Future[1] to "continue to assess the existing and potential risks associated with the military applications of artificial intelligence (AI) and the possible opportunities throughout their life cycle, in consultation with relevant stakeholders", this submission is intended to support States in ensuring that military applications of AI comply with existing legal frameworks and, where necessary, in identifying areas where additional legal, policy or operational measures may be required.

1. NORMATIVE PROPOSALS: REAFFIRMING EXISTING IHL AS THE STARTING POINT

The ICRC welcomes the strong support expressed by States, including in Resolution 79/239, for the application of international law – including IHL – to the use of AI in the military domain and the need to ensure that it is only used in compliance with this existing legal framework. IHL provides essential principles and rules that regulate the means and methods of warfare, including emerging military applications of AI, to protect those affected by armed conflict.

[1] UN General Assembly Resolution 79/1.

The ICRC has consistently emphasized that, while IHL does not explicitly prohibit or regulate the use of AI in military applications, it does restrict its development and use and places strict constraints on AI when it is integrated into weapon systems or used in some way to conduct warfare.[2] This already includes certain redlines, such as the prohibition of AI-enabled indiscriminate weapons, and AI-enabled biological and chemical weapons.

In addition, the requirement for States to conduct legal reviews under IHL applies to both AI-enabled systems that are integrated into weapons (the "means" of warfare) and those that influence how weapon systems are used (the "method" of warfare).[3] States should also conduct legal reviews of AI systems used in military decision-making where they affect the employment of weapon systems, enable cyber operations or are used as part of detention operations. These legal reviews serve as a key mechanism to ensure compliance with IHL, particularly as AI technologies become more integrated into military decision-making.

Recommendations

Existing and emerging normative proposals on the military application of AI should build upon established international legal frameworks and mechanisms, including IHL. Where necessary, these frameworks can be reinforced through the development of additional legal instruments, operational guidance or policy measures to address specific risks or challenges posed by emerging technologies. The form and content of such measures may vary depending on the specific use case. The ICRC encourages the international community to engage in concrete discussions on particular applications of AI in the military domain and to prioritize consideration of those that pose the greatest risks to people affected by armed conflicts.

2. A HUMAN-CENTRED APPROACH TO MILITARY AI

The ICRC welcomes the resolution's clear recognition of the need for a human-centric approach to the use of AI in the military domain. The ICRC has similarly advocated for a human-centred approach to the development and use of AI in armed conflict.[4] This approach has at least two key dimensions: first, ensuring a focus on the humans who may be affected by the use of AI; and second, emphasizing the obligations and responsibilities of the humans using or ordering the use of AI in military operations.

Despite the growing development of AI-related technologies in the military domain, it is important to recall that IHL requires individuals to make legal determinations. Humans must, for instance, determine the lawfulness of attacks that they plan, decide upon or execute, and they remain accountable for those determinations. The ICRC considers that human judgement is crucial for reducing humanitarian risks, addressing ethical concerns and ensuring compliance with IHL. Accordingly, when discussing military applications of AI, it is important to emphasize that while certain technical tasks may be carried out by machine processes, it is not the system itself that must comply with the law, but the humans using it.[5]

This does not mean that commanders and combatants cannot or should not use tools, including AI-decision-support systems (AI-DSS). Provided they are appropriately designed and used, AI-DSS may help to avoid or reduce civilian harm (see Section 3B below). However, these computer outputs and tools must only be designed and used to support, rather than hinder or replace, human decision-making.[6]

[2] This has also been affirmed by States, including in the UN General Assembly with Resolution 79/239.
[3] Protocol I additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts, 8 June 1977 (AP I), Article 36.
[4] ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021).
[5] ICRC, International Humanitarian Law and the Challenges of Contemporary Armed Conflicts: Building a Culture of Compliance for IHL to Protect Humanity in Today's and Future Conflicts (IHL Challenges Report), 2024, p. 61.
[6] *Ibid*. See also ICRC, IHL Challenges Report, 2019, p. 32.

Recommendation

In all considerations related to the development and use of AI in military applications, States and parties to armed conflicts must ensure that human control and judgement are preserved in decisions that pose risks to the life and dignity of people affected by armed conflict. This is essential for ensuring respect for applicable laws, including IHL, and upholding ethical standards.[7]

3. SPECIFIC APPLICATIONS OF AI IN THE MILITARY DOMAIN

The ICRC has identified three specific applications of AI in the military domain that pose particularly significant risks to people affected by armed conflict:

(A) AI in Autonomous Weapon Systems

Resolution 79/239 acknowledges the increasing integration of AI into weapons and weapon systems, a development that raises significant legal and humanitarian concerns. This includes weapons that may already constitute autonomous weapons or could become such weapons through software updates or changes in military doctrine.

The integration of AI, particularly machine learning (ML) techniques, into autonomous weapon systems (AWS) exacerbates existing challenges posed by AWS in ensuring compliance with IHL. In particular, it increases difficulties for human users to understand, predict, and control the system's functioning and effects. Users of AWS must be able to, with a reasonable degree of certainty, predict the effects of that weapon in order to determine whether it can be directed at a specific military objective, and take steps to limit those predicted effects, as required by IHL. This entails the ability to understand the functioning of the AWS: the nature and functioning of its sensors, the definition of its target profile and the potential effects in the circumstances of its use, including any risk of error or malfunction. This prohibition will be particularly relevant for AWS that function in opaque ways (the "black box" challenge), such as AWS relying on AI techniques, which prevent the human user from being able to understand, predict or explain the system's output. This impossibility effectively results in a lack of control over the weapon's effects, rendering it indiscriminate by nature.

This concern would also arise in AWS that incorporate ML, the functioning of which may change after an attack commences, resulting in force that may be applied in circumstances and in a manner unforeseen to the human user. Complex swarm technologies may also exhibit emergent behaviours that cannot be adequately understood, predicted or explained.

In this regard, we reiterate the joint call made by the ICRC President, with the UN secretary-general,[8] for new, legally binding rules prohibiting certain AWS and constraining the use of others.[9] In particular, we recommend a prohibition on unpredictable AWS – those that, due to their design or the circumstances and manner of use, do not allow a human user to understand, predict and explain the system's functioning and effects.

The ICRC supports all efforts by States to urgently adopt a legally binding instrument to regulate AWS, in whichever forum they choose.[10] We therefore continue to support the work of the Convention on Conventional Weapons Group of Governmental Experts. The integration of AI into AWS should also be considered when discussing normative proposals on military applications of AI. Doing so is essential to ensure a consistent and comprehensive approach to the regulation of military AI, to avoid normative gaps, and to effectively address the serious legal, ethical, and humanitarian risks that are exacerbated by the integration of AI into AWS. In this regard, the ICRC considers it important that binding prohibitions and restrictions on AWS, including AWS that incorporate AI, are integrated into broader discussions on the governance of military AI.

---

[7] ICRC, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, 2024 p. 8.
[8] ICRC, Joint call by the United Nations Secretary-General and the President of the International Committee of the Red Cross for States to establish new prohibitions and restrictions on Autonomous Weapon Systems, 2023.
[9] ICRC, ICRC Submission on AWS to the UN Secretary-General, 2024, p. 6.
[10] *Ibid*.

Recommendations

In light of the serious risks of harm to the people affected by armed conflict, challenges for compliance with IHL and ethical concerns raised by AWS, the ICRC has, since 2021, been calling for new, binding international rules on the development and use of AWS.[11] These rules should clarify and formalize specific prohibitions and restrictions concerning the design and use of AWS. Any such limits would be additional and complementary to existing IHL rules, including weapons treaties, and would not displace them. They would strengthen and build on existing legal protections in order to respond to the specific risks and ethical concerns raised by AWS. In particular, new rules must:

- prohibit unpredictable autonomous weapons that do not allow a human user to understand, explain or predict the system's functioning and effects;
- prohibit autonomous weapons designed or used to target humans directly. This is required because of the significant risk of IHL violations and the unacceptability of anti-personnel autonomous weapons from an ethical perspective.[12]

Even in the case of an AWS that is sufficiently predictable, and designed and used only to target objects, the user's reduced ability to know all the specifics of an attack, including the ultimate target and any incidental harm, will still create residual challenges for the context-specific application of IHL rules on the conduct of hostilities. To reduce the risk of violations, new rules must also strictly constrain the design and use of AWS, including through a combination of:

- restricting targets of the AWS to only those that are military objectives by nature;
- limiting the duration and geographic scope of the operation of the AWS;
- limiting the scale of use, including the number of engagements that the AWS can undertake;
- limiting the situations of use, namely constraining them to situations where civilians or civilian objects are not present;
- ensuring, to the maximum extent feasible, the ability for a human user:
    - to effectively supervise; and
    - in a timely manner, to intervene and, where appropriate, deactivate operation of the AWS;[13]
    - where this is not feasible, equipping the AWS with an effective mechanism for self-destruction or self-neutralization, which is designed so that the AWS will no longer function as an AWS when it no longer serves the military purpose for which it was launched.[14]

Against the backdrop of rapid and expanding development and use of AWS, the establishment of these prohibitions and restrictions on AWS, in clear and binding international law, is an urgent humanitarian priority. In 2024, the ICRC submitted its views in more detail, for consideration by States and the UN secretary-general, on how these rules could be drafted in a legally binding instrument.[15]

(B) AI in military decision-making

As highlighted in the secretary-general's report on "Current developments in science and technology and their potential impact on international security and disarmament efforts" (A/79/224), one of the most widespread military applications of AI is in systems intended to support military decision-making.[16] Commonly referred to as AI-DSS, these computerized tools bring together data sources –

[11] ICRC, ICRC Position on autonomous weapon systems, 2021.
[12] ICRC, ICRC Submission on AWS to the UN Secretary-General, 2024, p. 6.
[13] *Ibid*.
[14] Language of the Convention on Certain Conventional Weapons, Amended Protocol II.
[15] ICRC, ICRC Submission on AWS to the UN Secretary-General, 2024.
[16] UNGA, Report of the Secretary-General: Current developments in science and technology and their potential impact on international security and disarmament efforts, 2024, para. 5.

such as satellite imagery, sensor data, social media feeds or mobile phone signals – and draw on them to present analyses, recommendations and predictions to decision makers.

One of the main military benefits of AI-DSS that is touted, and is behind their development and use, is their ability to accelerate planning and decision-making processes, giving an advantage over the adversary. Increasing the speed of military operations can, however, create additional risks for both civilians and combatants, including by increasing the risks of miscalculation and escalation.[17]

Integrating AI into decision-support systems also raises a number of concerns related to system functioning, data quality and human-machine interaction. One concern is the potential for these systems to increase the rate of unforeseen errors, and to perpetuate or amplify problematic biases — particularly those based on age, gender, ethnicity, or disability. Research indicates that these challenges will increase with more complex forms of AI, such as ML, which can make it more difficult, or even impossible, for the user to understand how and why the system generates its output from a given input. Moreover, in some cases, when a number of different decision-support systems build on and contribute to decisions in a single process, an error in one system can compound or cascade errors across an entire planning and decision-making process.[18]

Generally, AI-based systems will perform better when given clear, well-defined goals and access to representative and high-quality data. However, armed conflict environments are marked by uncertainty and volatility, as well as deliberate deception techniques by adversaries. These features make it extremely difficult to obtain reliable or transferable data. Even where good data exists, it may not reflect the specific operational or humanitarian dynamics of a particular context.[19] Moreover, for AI systems that rely on training data, the utility of those data can rapidly diminish once a conflict begins. Parties to armed conflicts will continuously seek to maintain the initiative and operate in a manner that is not anticipated by their adversary, adapting their strategies and tactics accordingly. This can fundamentally alter the environment in which the system was expected to operate, making the original data no longer representative of the new operational conditions. In such cases, the system's outputs may become unreliable, and the AI model may require re-evaluation or retraining in order to remain fit for purpose.

Human interaction with these systems raises further concerns. Users may exhibit a cognitive tendency known as "automation bias", which is a propensity to rely on machine outputs even when other available information may call those outputs into question. This bias is particularly pronounced in high-pressure or stressful environments, such as those typical of armed conflicts.[20]

Taken together, these factors can hamper a user's ability to scrutinize the information available. The practical consequence might be, for instance, that someone plans, decides upon or launches an attack based solely on an AI-DSS's output, thereby effectively serving as a human rubber stamp rather than assessing the lawfulness of the attack by considering all the information reasonably available including the AI-DSS output.[21]

As outlined above, determinations on the lawfulness of an attack – including whether or not an object is a military objective – must be made by a human.[22] This is not to say that, in making these legal assessments, commanders and combatants cannot, or even that they should not, use tools – including AI-DSS. Indeed, the careful use of AI-based systems may facilitate quicker and more comprehensive information analysis, which can support decisions in a way that enhances IHL compliance and

[17] ICRC, IHL Challenges Report, 2024, p. 66.
[18] *Ibid*., pp. 64-65; ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021). See also ICRC, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, 2024, p. 31 and p. 54.
[19] *Ibid.*
[20] ICRC and the Geneva Academy, Expert Consultation report – Artificial Intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed Conflicts: Current Developments and Potential Implications, ICRC, 2024, p. 17.
[21] ICRC, IHL Challenges Report, 2024, p. 65.
[22] *Ibid.*, p. 61.

minimizes risks for civilians. In fact, States have already adopted a broad range of military decision-making tools, at all levels, to assist members of their armed forces during the planning, ordering and conduct of attacks. In some States, for instance, the operational process of estimating incidental civilian casualties is computerized, and used by commanders as one source of information in their assessment of whether an attack will be proportionate under IHL. In the context of urban warfare in particular, the ICRC has recommended that online open-source repositories should be used to gather information about the presence of civilians and civilian objects.[23] AI tools can likely assist in collecting and synthesizing such sources. The use of AI-DSS to support weaponeering may also inform the choice of means and methods of attack that can best avoid, or at least minimize, incidental civilian harm.[24]

Importantly, IHL imposes obligations to take constant care to spare the civilian population and to take all feasible precautions in attack. Therefore, in developing and using AI-DSS, armed forces should be considering not only how such tools can assist them to achieve military objectives with less civilian harm, but also how they might be designed and used specifically to protect civilians. This could include tools to recognize and track civilian populations and alert forces to their presence, or to recognize distinctive emblems or signals that indicate protected status. However, the important point is that these computer outputs can inform but must not displace the need for legal determinations. In the ICRC's view, this means that in designing and using any AI-DSS, militaries and other armed actors must account for the ways in which these AI tools function and the tendencies of human users interacting with them.[25]

In addition to its use in targeting decisions, militaries are also exploring the use of AI to support other operations traditionally carried out by humans, including detention operations. Future detention operations are likely to involve the use of AI to support decisions on who should be detained, as well as the management of detention facilities.[26] While technology deployed responsibly and with robust human oversight can contribute to IHL compliance, it also carries risks including bias, lack of transparency, and faulty programming and analysis, all of which can undermine humane treatment and compliance with IHL. In addition, if authorities step away from direct human contact with detainees, they will also give up critical insights required for taking well-informed and timely decisions. As such, when using AI-DSS in detention operations, States must ensure that it does not adversely affect the treatment of detainees and the conditions of detention. Detention authorities must notably retain direct contact with the detainees, which is essential to build trust, foster situational awareness, maintain order without force, identify problems early, and ensure that detention conditions comply with IHL.[27]

Recommendations

To support efforts by States and other actors to ensure that military uses of AI-DSS remain consistent with IHL and humanitarian principles, the ICRC has formulated a non-exhaustive set of **preliminary recommendations** (see Annex) relating to the development and use of AI-DSS in armed conflict. It has done so by drawing upon its own research and discussions with experts, its extensive experience in providing humanitarian assistance to people affected by armed conflict worldwide, as well as its mandate to prevent suffering by promoting and strengthening IHL.[28]

These recommendations reflect the ICRC's human-centred approach to the development and use of AI in armed conflict (see Section 2 above). They aim to ensure that AI-DSS can support, rather than undermine, human judgement, legal compliance and the protection of those affected by armed

[23] *Ibid.*, p. 66; ICRC, Reducing Civilian Harm in Urban Warfare: A Handbook for Armed Groups, 2023, p. 15.
[24] AP I, article 57(2)(a)(ii); ICRC, Customary IHL Study, Rule 17.
[25] ICRC, IHL Challenges Report, 2024, p. 64.
[26] *Ibid.*, p. 22.
[27] *Ibid.*
[28] ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021); ICRC observations on expert report: 'Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making,' 2024; ICRC, IHL Challenges Report, 2024, p. 61; ICRC and the Geneva Academy, Expert Consultation report – Artificial intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed conflicts: Current Developments and Potential Implications, ICRC, 2024.

conflict. They focus on 1) ensuring human control and judgement; 2) system design requirements; 3) testing, evaluation, verification and validation; 4) legal reviews; 5) operational constraints on use; 6) user training; 7) after-action reviews; and 8) accountability, among others. The recommendations are in the **Annex** to this submission.

(C) AI in Information and Communications Technologies

AI is expected to change how actors defend against and conduct information and communications technology (ICT) activities, including in armed conflict. In particular, States have noted with concern that the use of AI and other emerging technologies in malicious ICT activities may further increase their scale and speed, as well as the harm they may cause.[29] For example, AI enables tools to identify and develop exploits for new vulnerabilities in software or networks, or to conduct harmful ICT activities autonomously, whether in offence or in defence. The ICRC is concerned that this could increase the risks of indiscriminate attacks, incidental civilian harm, including damage to critical civilian infrastructure, as well as the uncontrolled escalation of conflict, particularly in complex and interconnected digital environments.[30]

Similarly, information or psychological operations are not a new feature of armed conflicts; however, AI is changing how information is created and spread. AI-enabled systems, particularly generative AI, have been widely used to produce harmful content – text, audio, photos and video – which is increasingly difficult to distinguish from authentic, original content.[31] The ICRC is concerned about the consequences for civilians that might result from the creation and spread of such information through ICT, including information that contributes to or encourages violence, causes lasting psychological harm, undermines access to essential services or disrupts the operations of humanitarian organizations.

Recommendations

In light of these concerns, the ICRC underlines the importance of applying existing international law, including IHL, to the use of AI in ICT activities. The ICRC urges States to ensure that the development and use of AI-supported ICT activities respect the protections afforded to civilians and civilian infrastructure in armed conflict. Moreover, in light of the emergence of increasingly autonomous ICT capabilities, the ICRC further encourages States to address the serious challenges posed by these tools, particularly by considering whether existing international law, including IHL, provides sufficient safeguards against the harm such tools can cause, or whether additional limits are needed.

4. CONCLUSION

The ICRC is grateful for the opportunity to share its above views and recommendations on ways to address the challenges and concerns raised by AI for the secretary-general's consideration, and stands ready to contribute further to assist States in taking effective action to address the risks posed by AI applications in the military domain.

11 April 2025

[29] 34th International Conference of the Red Cross and Red Crescent, Resolution 2 "Protecting civilians and other protected persons and objects against the potential human cost of ICT activities during armed conflict", 2024.
[30] ICRC, IHL Challenges Report, 2024, pp. 66-67.
[31] *Ibid*., pp. 58-59.

ANNEX: PRELIMINARY RECOMMENDATIONS ON MILITARY ARTIFICIAL INTELLIGENCE DECISION-SUPPORT SYSTEMS

Armed forces are investing heavily in military applications of artificial intelligence (AI) and related data-collection and analysis technologies, which have significant implications in armed conflict. Some of the most widespread and increasingly prominent military applications of AI relate to military decision-making, particularly the use of these technologies as part of decision-support systems (DSS). These computerized tools bring together data sources – such as satellite imagery, sensor data, social media feeds or mobile phone signals – and provide outputs (e.g. analyses, recommendations and predictions) to inform human decision-making. The possible uses of DSS in the military domain are broad, from supporting decisions about who, what or where to attack, to decisions on who to detain and for how long. They can also provide recommendations on military strategy and specific operations, particularly attempts to predict or pre-empt adversaries' actions,[32] and give recommendations on troop movements and logistics, such as supply chain management and equipment maintenance.

The International Committee of the Red Cross (ICRC) has consistently emphasized that while international humanitarian law (IHL) does not explicitly prohibit the military application of AI, it does apply to limit its development and use by parties to armed conflicts. IHL also imposes strict constraints on AI when it is integrated into weapon systems or used in other ways to conduct warfare,[33] with the aim of protecting people affected by armed conflicts from the dangers arising from military operations. This already includes certain redlines, for example, existing prohibitions of weapons, such as indiscriminate weapons or chemical weapons that would of course also be prohibited if operated with the support of AI.

However, the use of AI-decision-support systems (AI-DSS) in military decision-making on the use of force raises particular challenges that require putting in place specific measures and operational constraints to reduce risks for people affected by armed conflicts, and to facilitate compliance with IHL. With this in mind, the ICRC has formulated a non-exhaustive set of preliminary recommendations relating to the development and use of AI-DSS on the use of force in armed conflict. It has done so by drawing upon its own research and discussions with experts, its extensive experience in providing humanitarian assistance to people affected by armed conflict worldwide, as well as its mandate to prevent suffering by promoting and strengthening IHL.[34]

These recommendations align with a 'human-centred approach' to the development and use of AI in armed conflict.[35] Such an approach focuses on: (1) the humans who may be affected, and (2) the obligations and responsibilities of the humans using or commanding the use of AI.

The recommendations also have the overall goal of preserving human judgement in military decision-making on the use of force in armed conflicts. The ICRC considers that human judgement is crucial to reducing humanitarian risks, addressing ethical concerns and ensuring compliance with IHL. In particular, IHL obligations regarding the conduct of hostilities must be fulfilled by human commanders and combatants, who are responsible for determining the lawfulness of the attacks they plan, decide upon or execute, and who remain accountable for these assessments. Likewise, decisions on whether and for how long a person may be deprived of liberty – including humanitarian considerations related to age, health, and other individual circumstances – can only properly be made by human authorities. When discussing AI-DSS and compliance with IHL, it is important to emphasize that it is not the system itself that must comply with the law, but the humans using it. As such, AI-DSS, including DSS based on machine learning, must remain tools that help and support human decision-making, rather than tools

[32] ICRC, Decisions, Decisions, Decisions: computation and Artificial Intelligence in military decision-making, 2024, p. 3.
[33] This has also been affirmed by states, including in the UN General Assembly with Resolution A/C.1/79/L.43.
[34] ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021); ICRC Observations on expert report: Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, 2024; ICRC, IHL Challenges Report, 2024, p. 61; ICRC and the Geneva Academy, Expert Consultation report – Artificial intelligence and Related Technologies in Military Decision-Making on the Use of Force in Armed conflicts: Current Developments and Potential Implications, 2024.
[35] ICRC, AI and machine learning in armed conflict: A human-centred approach, 2019 (updated in 2021).

that impair or displace it. In the ICRC's view, this means that in designing and using any AI-DSS, militaries and other armed actors must account for the ways in which these AI tools function, and the cognitive and behavioural challenges that may arise when humans interact with them.

These preliminary recommendations are variously directed at the "users" (including military decision-makers and others commanding the use of these systems), and "developers" (including all those engaged in their design, development, or programming of these systems). They focus on 1) ensuring human control and judgement; 2) system design requirements; 3) testing, evaluation, verification and validation; 4) legal reviews; 5) operational constraints on use; 6) user training; 7) after-action reviews; and 8) accountability, among others.

Guiding Principle: Preserving human control and judgement

1. Human judgement and control: In all considerations related to the development and use of AI-DSS in armed conflicts, States and parties to armed conflicts must ensure that human control and judgement are preserved in decisions that pose risks to the life, liberty, and dignity of people affected by armed conflict. AI-DSS, including DSS based on machine learning, must remain tools that help and support, rather than impair or displace, human decision-making. This is essential for ensuring respect for applicable laws, including IHL, and upholding ethical standards.[36]

Design, development and review

2. Interfaces and platforms: Developers should ensure that AI-DSS user interfaces and platforms are designed in such a way that (1) they present users with information in a coherent way, without overwhelming them to the extent that they can no longer meaningfully engage with the information, and (2) they allow users to challenge the output. This means designing these systems to maximize transparency and explainability in how the system functions, including enabling users to understand the causal link between inputs and outputs, the functioning of algorithmic processes, and the weighting given to and the currency of underlying datasets.

3. Datasets: Developers and users should ensure that the datasets used as input for AI-DSS are reliable, ethically sourced and validated, as well as representative of the operating context, including the human terrain and civilian environment in which the AI-DSS may be expected to be used. This should include data on civilian presence, infrastructure, movement patterns and behaviours, which are critical for distinguishing military objectives from civilian objects and for ensuring compliance with IHL (see recommendation 5 below on testing and validation). This may require data from multiple sources, including intelligence, surveillance and reconnaissance (ISR) systems, historical combat data and military simulations.[37]

4. Mitigating bias to prevent discriminatory outcomes: Developers and users should take measures to mitigate gender, racial, ethnic, disability and other similar forms of bias in the design and use of AI-DSS, including in the underlying datasets and training methods. These measures should take account of information regarding, gender, race, and other relevant criteria characterizing the population. They should include, where feasible and appropriate, data disaggregated by sex, age and disability; enrich civilian-related data to capture potentially relevant factors; ensure that datasets used are context-appropriate; and involve advisers with multidisciplinary expertise and diverse backgrounds in the design, development and use of AI-DSS.

5. Testing and validation: Developers should ensure that AI-DSS undergo rigorous testing and validation in environments that simulate the complexity of armed conflict, including not only the rapidly evolving operational environment and likely conduct and deception strategies of an adversary, but also realistic civilian presence, activities, actions and reactions. This testing should consider how users are likely to interact with the AI-DSS, and any impact the system may have on their ability to comply with IHL, especially the requirement to identify and distinguish between military objectives and civilian objects, and between combatants and civilians or other protected persons, as well as

[36] ICRC, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in military decision-making, 2024, p. 8.
[37] ICRC, IHL Challenges Report, 2024, p. 66.

decisions related to detention and internment. This includes assessing risks such as automation bias, where the users may over-rely on system outputs, and implementing measures to minimize such risks. Additionally, developers should apply particularly stringent reliability and safety assurance standards given the gravity of risks posed by erroneous AI-DSS outcomes, especially in decisions on the use of force. They should also ensure that AI-DSS produce outputs that are sufficiently predictable and reliable to ensure the system can actually support rather than hinder the user's decision-making process.

6. Re-testing: Developers should ensure an AI-DSS is re-tested whenever its intended use changes, or if it is modified in a manner that alters its function or effects. Additionally, developers should ensure re-testing when after-action reviews, battle damage assessments or lessons learned from actual use in conflict indicate that the system's outcomes do not meet expectations. Limitations of testing and validation should be recorded and made available for users, including when such systems have only been tested in simulation conditions.

7. Legal reviews: States must include AI-DSS in legal reviews where they form part of weapon systems or form part of the process in which a weapon system is designed or expected to be used.[38] These reviews must take into consideration the technical features of AI-DSS, and the cognitive and behavioural tendencies of users interacting with them, such as automation bias, as well as the particular challenges of their use in armed conflicts. Such reviews must be updated whenever an AI-DSS is used in a new way or is modified in a manner that alters its function or effects. More broadly, States should consider conducting these kinds of reviews of all AI-DSS involved in military decisions to use force, detain, or intern in armed conflict, at a minimum as a matter of policy or good practice.[39]

8. AI-DSS for enhancing civilian protection: States and developers should consider not only how AI-DSS can assist armed forces to meet their objectives in attacks and other operations, but also how AI-DSS can be specifically designed and used to facilitate better compliance with IHL[40] and increase protection for civilians. This includes prioritising research and investment in identifying, or building and maintaining, datasets that can support assessments of, the "human terrain,"[41] including how civilians may react to or otherwise be impacted by military operations.

Use of AI-DSS

9. Restricting use cases: Before determining how an AI-DSS should be used for military purposes, users should first assess whether such a tool should be used at all. This requires evaluating the system's suitability for the specific task and context, ensuring it has clear, well-defined goals and access to high quality data.[42] In some cases, AI-DSS may need to be excluded altogether from certain use cases in armed conflict. One clear example would be that such tools must never be incorporated into nuclear-weapon command-and-control systems.[43] Furthermore, and in line with the aforementioned recommendations on system testing and legal reviews, armed forces should restrict the use of AI-DSS to tasks and contexts for which they have been specifically and rigorously tested and reviewed.

10. Ruling out the use of AI-DSS in conjunction with autonomous anti-personnel targeting: AI-DSS must not be used to generate target profiles or recommendations that feed directly into weapon systems that are designed or used to target human beings autonomously, i.e. without further human intervention.[44]

---

[38] See article 36, AP I.
[39] See article 1 common to the four 1949 Geneva Conventions.
[40] Including so-called 'positive' obligations such as the constant care obligation, or the obligation to account for persons reported missing in armed conflict.
[41] Encompassing information on population demography, locations, density, humanitarian situation and civilians' likely courses of action before and during conflict: see further ICRC, Reducing Civilian Harm in Urban Warfare: A Commander's Handbook, 2023, p. 38.
[42] ICRC, IHL Challenges Report, 2024, p. 65.
[43] *Ibid*.
[44] *Ibid*., p. 62; ICRC, ICRC Submission on AWS to the UN Secretary-General, 2024, pp. 5-6.

11. Ensuring lawful and responsible use: For use cases where an AI-DSS is not prohibited, States need to develop and adapt appropriate doctrinal frameworks, standard operating procedures (SOPs), tactics, techniques and procedures (TTPs) and other guidance to ensure compliance with IHL and other applicable laws and policies. These frameworks should account for the specific risks associated with AI-DSS and prevent its use in ways that could exacerbate unlawful behavior, increase risk of IHL violations, or heighten the risk of harm to civilians.

12. User training: Armed forces should ensure that any user of an AI-DSS receives specific training to develop an adequate understanding of how the system functions, how it generates outputs and its limitations and vulnerabilities (e.g. vulnerabilities to adversarial attacks). Training should also equip users to recognise and mitigate the effects of human cognitive tendencies (e.g. automation bias) that may arise when interacting with such systems. This should be supplemented by ongoing training, exercises, and simulations in realistic settings.

13. Assessing information from all sources reasonably available: When drawing on an AI-DSS output, in particular for decisions on targeting, users must assess information from all sources reasonably available. Provided they are appropriately designed and used, AI-DSS may offer benefits in this respect. However, AI-DSS outputs must not be relied upon as the sole basis for information to make legal assessments, as information from other sources may identify situations where AI-DSS outputs are incomplete, inaccurate or fail to capture context-specific factors. Users of AI-DSS need to cross-check the outputs of these tools against all other reasonably available intelligence, whether it is for targeting decisions or for decisions on who to intern or detain.

14. Use of AI-DSS in detention operations in armed conflict: When using AI-DSS in detention operations, States must ensure that it does not adversely affect the treatment of detainees and the conditions of detention or internment. Detaining authorities should notably retain direct contact with the detainees, which is essential to build trust, foster situational awareness, maintain order without force, identify problems early, and ensure that detention conditions comply with IHL.[45]

15. Tactical patience: Armed forces should consider how to maintain long-employed practices such as tactical patience when employing AI-DSS. This may involve purposely slowing down at certain points in planning and decision-making processes to preserve time for deliberation and careful consideration of decisions on the conduct of hostilities in armed conflict.

16. After-action reviews: Users should conduct timely and objective after-action reviews of use cases involving AI-DSS, examining the reliability of the AI-DSS output in context, and how users interacted with and responded to the AI-DSS generated outputs and prompts. These reviews should assess, *inter alia*, whether the AI-DSS functioned as intended, and any impact its use had on civilians, civilian objects or parts of the natural environment, or on the user's ability to comply with IHL. A process should be in place to suspend or discontinue the use of AI-DSS when issues affecting reliability, compliance, or safety are identified until they have been resolved by modifying, re-testing, and re-reviewing it before any further use.

[45] ICRC, IHL Challenges Report, 2024, p. 22.