HUMANITARIAN LAW & POLICY


# ‘Constant care’ must be taken to address bias in military AI

*August 28, 2025*, Accountability / Analysis / Artificial intelligence in military decision-making / Generating Respect for IHL / IHL / New Technologies / Special Themes

11 mins read

**Laura Bruun**
Researcher, Stockholm International Peace Research Institute (SIPRI)

**Marta Bo**
Associate Senior Researcher, SIPRI

*As many states, especially those with large and resourceful militaries, are exploring the potential of using artificial intelligence (AI) in targeting decisions, there is an urgent need to understand the risks associated with these systems, one being the risks of bias. However, while concerns about bias are often mentioned in the military AI policy debate, how it manifests as harm and what can be done to address it is rarely discussed in depth. This represents a critical gap in efforts to ensure the lawful use of military AI.*

*To help bridge this gap, Laura Bruun and Marta Bo from the Stockholm International Peace Research Institute (SIPRI) unpack the humanitarian and legal implications of bias in military AI. They show how bias in military AI is likely to manifest in more complex and subtle ways than portrayed in policy debates, and if unaddressed, it may affect compliance with IHL principles of distinction, proportionality, and, especially, precautions in attack.*

*ICRC Humanitarian Law & Policy Blog · 'Constant care' must be taken to address bias in military AI*

The use of artificial intelligence (AI) in targeting decisions in warfare is no longer a future scenario. While increased speed and scale are often highlighted as drivers for using AI in targeting decisions, speed and scale also **compound** the risks associated with military AI, one of which is bias.

Bias in military AI *refers* to the phenomenon of a AI system – often unintentionally – being inclined towards or against certain individuals or groups of people in a way that is systemic and unfair. A military AI system may contain bias if it is trained on datasets that over- or under-represent certain ages, dialects, genders or skin colours. Bias in military AI takes many shapes, but the concern is the same: it increases the risks of unintended harm. As more states are showing interest in using military AI to inform, support and even execute targeting decisions, there is an urgent need to understand – and address – the risks of bias.

## Old problem, new dimensions

Bias is not a new problem in military settings. Throughout history, military organizations and personnel have made decisions with dire humanitarian consequences due to unconscious or subconscious assumptions and inclinations (see, for example, *here* and *here*). However, the use of AI in military decision-making exacerbates existing concerns related to bias. First, the design and programming of an AI system inevitably *reflect* biases of society, developers and users. Second, the increased speed, scale and opacity of these systems will likely amplify the risk of harmful bias manifesting. Their use accelerates the potential for flawed target recommendations and leaves less, if any, time for users to spot and correct potential biases.

States involved in military AI policy processes are increasingly raising concerns about bias (see for example *here*, *here* and *here*), often highlighting gender and racial bias. That said, states have only scratched the surface of the issue: how bias in military applications of AI manifest as harm and what concretely can be done to address has so far been relatively absent from discussions. To support efforts to understand – and thus better mitigate – bias in military AI, SIPRI is this week publishing a report exploring its humanitarian and legal implications. The report is informed by expert interviews and an in-person workshop with participants from, among other, governments, industry, and civil society, and its main findings are as follows:

## Bias in AI could lead to civilians and civilian objects being seen as threats – or not seen at all

Bias in military AI often manifests in more complex and subtle ways than portrayed in policy debates. Beyond gender and racial bias, military AI systems are likely to contain biases around, for example, disability status, age, socio-economic class, language, and culture. If these biases go unaddressed, they may undermine compliance with key IHL principles of *distinction*, *proportionality*, and, especially, *precautions in attack*. The ways in which bias in military AI can lead to harmful, and potentially unlawful, outcomes, can be boiled down to two main risk pathways:

*Risk pathway #1: Bias in military AI may lead to misidentifying non-threats as threats*

Military AI systems used for target identification may – due to biased assumptions about enemy traits embedded in the AI – misidentify civilians and or other protected persons and objects as threats. For example, an AI system might have been programmed on the assumption that the use of prepaid phone cards is an indicator of enemy status, without taking into account the high use of prepaid phone cards among sub-sections of a population, such as migrant workers. Or it might have been trained to classify groups of armed men as a target, without considering that the possession of a rifle might indicate hunting or festive activities specific to certain cultures. The risk of bias leading to misidentifications is especially acute if using AI to identify fighters who are not uniformed. Instead of relying on clear visual cues like uniforms and enemy symbols as indicators for combat status, target profiles will need to be developed based on assumptions about hostile behavior, functions and characteristics.

Moreover, bias in military AI could also lead to misidentifying civilian objects. This is especially so if AI is used in attacks against military objects whose status as lawful targets are not military by nature but depends on the context, including human activity in and around such objects. For example, if an AI system used to identify temporary command centers is not trained on datasets that sufficiently reflect local religious or cultural practices, it may misinterpret peaceful gatherings as enemy activity.

One distinction is worth making: In the case of autonomous weapon systems (AWS), where the system both selects *and* engages targets, misidentifications would result in *direct* harm. In the case of AI-decision support systems (AI-DSS), the risk of harm is *indirect* – manifesting only if users act on the flawed target recommendations (as elaborated on *here*).

*Risk pathway #2: Bias in military AI may lead to civilians and civilian objects not being detected at all*

Bias in military AI may also lead to *failures* to detect civilians and civilian objects in the target area. This risk is particularly relevant if using AI to detect, predict and/or assess civilian presence in the vicinity of an attack. Civilian populations are *not a homogenous group*: they vary in terms of ethnicity, socio-economic class, cultural practice, language or disability status. However, if an AI system used to detect civilians is not trained on a dataset that sufficiently reflects the complexity and diversity of the local population, the system may fail to detect people or objects that look, move or act in ways not represented in the dataset. For example, if people in wheelchairs are underrepresented in the dataset, an AI-DSS used to assess the presence of civilians in a target area may simply fail to detect this segment of the population because it is not trained to do so, ultimately exposing them to greater harm if their presence is not accounted for.

## If not spotted, bias in military AI could lead to violations of IHL

Failures to account for bias in military AI could undermine compliance with IHL, particularly the principles of distinction, proportionality and precautions in attack. Reliance on AI systems for target identification risk violating the principle of distinction and precautions in attack insofar as militaries rely on *overly* generalized target profiles without doing everything feasible to verify that the objectives to be attacked are military objectives. Likewise, failing to account for all segments of the civilian population in an attack increases the risk that certain groups will be exposed to greater harm, potentially violating the principle of proportionality. It could also amount to a violation of the principle of precautions in attack, which requires parties to take *constant care* in military operations to spare civilians. Arguably, this obligation

requires parties to the conflict to take into account the diversity of the *whole* civilian population, including differences in gender, age, ethnicity, and disability, to ensure that all groups receive equal protection (see *here* and *here*).

Bias in military AI could amount to a violation of the *IHL prohibition against adverse distinction*, which prohibits discrimination based on, for example, race, religion and sex. However, while it is well-established that this prohibition applies to the treatment of persons in the power of a party, it is less clear whether it also applies to targeting (according to *some commentators*, it does). The possible application of this prohibition could guide states in meeting their obligations related to distinction, proportionality, and precautions in attack, particularly when using military AI.

## Addressing bias in military AI may not just be 'good practice'—it may be a legal obligation

Rather than aiming to eliminate all biases in military AI systems (a nearly impossible task) efforts can – and should – be taken to *mitigate* its risks. And in fact, measures to address bias may not just be a matter of voluntary 'good practice', but a legal requirement flowing from IHL's positive obligations to minimize harm to civilians. How bias mitigation measures translate into specific requirements and limits in the development and use of military AI systems remain to be explored by states and experts. However, key elements revolve around representative dataset and human control and judgment:

- Representative dataset: the more representative the underlying datasets are of the operational environment, the lower the risk is that the AI system will lead to misidentifications of threats or failures to spot protected people and objects. Ensuring that datasets sufficiently represent the environment of use requires developers of military AI systems to consider the intended use already in the data collection and processing phases, and it would require diversity in the teams that collect and label the data. However, a major flipside associated with more representative datasets is that the more context-specific the data is, the greater the need to collect information about local populations, potentially expanding surveillance measures and infringing on privacy rights.
- Human control and judgment: Bias cannot only be addressed through technical means; human control and judgment during use are key measures to spot, mitigate and prevent harmful bias in military AI. In the case of AI-DSS, users would need to have multiple layers of verification in place to ensure they can spot and correct biased AI outputs. Meanwhile, for AI-enabled AWS, where users may not be able to verify outputs once the system is activated, control and judgment over the systems' behavior and effects would need to be exercised in advance. This means limiting their use to situations where the risk of bias causing harm is low, i.e. to situations that do not depend on assessments of human activity and characteristics. An example of a use case with a low risk of bias manifesting as harm would be attacks against military objects that are clearly identifiable and located in areas without civilians.

## Bias mitigation should be an integral part of policy efforts to ensure IHL compliance in the development and use of military AI

The humanitarian risks flowing from bias in military AI are many and complex. Without a deeper understanding of what bias is and how it manifests, states claimed desire to reduce the risks will fall short. In fact, failures to address bias in military AI could amount to violations of IHL, notably obligations to take constant care in military operations to spare civilians. Addressing bias in military AI should, therefore, not be treated as a fringe concern – it should be an integral part of establishing what limits and requirements IHL places on military AI.

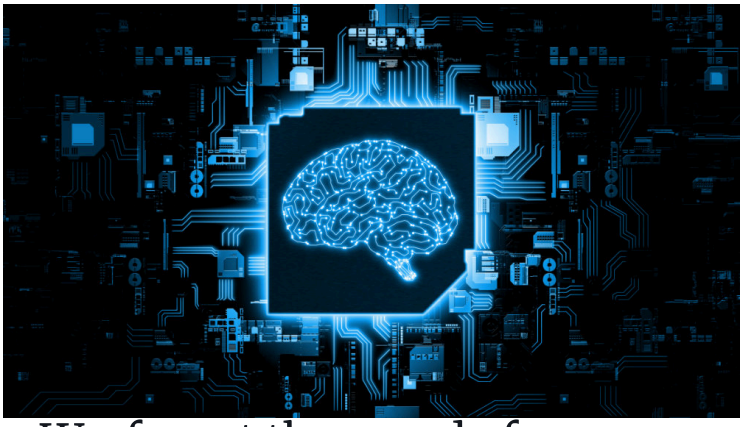
**Author's note**: *This post is based on a new SIPRI publication 'Bias in Military Artificial Intelligence and Compliance with International Humanitarian Law'*

## See also:

- Maria Haas, *How AI learns, and what it misses: why data selection matters in humanitarian action*, August 14, 2025.
- Joanna Wilson, *AI, war and (in)humanity: the role of human emotions in military decision-making*, February 20, 2025.
- Elke Schwarz, *The (im)possibility of responsible military AI governance*, December 12, 2024.
- Erica Harper, *Will AI fundamentally alter how wars are initiated, fought and concluded?*, September 26, 2024.
- Matthias Klaus, *Transcending weapon systems: the ethical challenges of AI in military decision support systems*, September 24, 2024.

Tags: AI, AI decision support systems, algorithmic bias, Artificial Intelligence, Autonomous Weapons, AWS, Civilians, IHL, international humanitarian law, Legal Review of Weapons, Means and Methods of Warfare, modern warfare, Respect for IHL

## *You may also be interested in:*



### Warfare at the speed of thought: can brain-computer interfaces comply with IHL?

13 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Generating Respect for IHL / IHL / New Technologies / Special Themes Anna M. Gielas

Brain-computer interfaces (BCIs) are no longer speculative technologies of future warfare – they are being ...

### How AI learns, and what it misses: why data selection matters in humanitarian action

16 mins read

Accountability / Analysis / Artificial intelligence in military decision-making / Generating Respect for IHL / IHL / New Technologies / Special Themes Maria Haas

Artificial intelligence (AI) is entering humanitarian action at pace, shaping decisions from crisis forecasting to ...