

HUMANITARIAN LAW & POLICY



Outsourcing humanity? International law, humane treatment, and artificial intelligence in detention operations

November 13, 2025, Accountability / Analysis / Detention / Generating Respect for IHL / IHL / Law and Conflict / New Technologies

🕒 14 mins read



Terry Hackett

Head, Persons Deprived of Liberty Unit, the International Committee of the Red Cross (ICRC)



Alexis Comminos

Thematic Legal Adviser, the International Committee of the Red Cross (ICRC)



As artificial intelligence (AI) begins to shape decisions about who is detained in armed conflict and how detention facilities are managed, questions once reserved for science fiction are now urgent matters of law and ethics. The drive to harness data and optimize efficiency risks displacing human judgment from one of the most sensitive areas of warfare: deprivation of liberty. In doing so, AI could strip detainees of what remains of their humanity, reducing them to data points and undermining the core humanitarian guarantees that the Geneva Conventions were designed to protect.

In this post, Terry Hackett, ICRC's Head of the Persons Deprived of Liberty Unit, and Alexis Comminos, ICRC's Thematic Legal Adviser, explore how the use of AI in detention operations intersects with international humanitarian law (IHL), and why humane treatment must remain a human-centered endeavor. Drawing on the ICRC's recent recommendations to the UN Secretary-General, they argue that while IHL does not oppose innovation, it sets the moral and legal boundaries that ensure technological progress does not come at the cost of human dignity.

ICRC Humanitarian Law & Policy Blog · Outsourcing humanity? International law, humane treatment, and artificial intelligence in detention operations

For people detained in connection with armed conflict, life depends entirely on the decisions of their captors. Cut off from family and community, they are uniquely vulnerable to risks of arbitrary detention, ill-treatment, disappearance, and neglect. When technology enters this fragile human sphere, the stakes rise even higher. The introduction of artificial intelligence into detention operations risks reducing individuals to data points, their fates decided by opaque systems that cannot feel pain, doubt, or compassion.

Recognizing this danger, the ICRC sounded the alarm in its [2024 IHL Challenges Report](#), warning that bias, lack of transparency, and diminished human oversight could erode compliance with international humanitarian law. The risk extends across the entire detention spectrum, from determining who should be interned, to managing facilities, to reviewing and releasing detainees.

State compliance with IHL must serve as a guardrail, ensuring that technology enhances rather than erodes humane treatment. The Geneva Conventions require^[1] state parties to respect and ensure respect for IHL, including the humane treatment of people deprived of liberty – an obligation that applies equally to AI-driven systems. The discussion that follows draws on the [ICRC's preliminary recommendations](#) to the United Nations Secretary-General, examining three dimensions essential to keeping humanity at the centre of technological progress in war: a human-centred approach, robust safeguards and testing, and ongoing legal review.

Keeping humans in the loop

AI, including predictive analytics, has the potential to support human decision making, improve services to detainees and manage places of detention more efficiently. However, it is ultimately how AI is designed, implemented and used by humans that will determine if humane treatment is preserved or degraded.

In the ICRC's view, IHL requires that legal determinations be made by humans, because it is humans – and the states or armed groups they represent – on whom the law imposes legal obligations. This core principle remains valid if and when humans use AI, including to generate recommendations that inform their decision-making.^[2]

It can be argued that AI has the potential to bring greater impartiality and consistency than humans may achieve. However, the presumption that AI is impartial by nature is fallacious: it may be conceived or trained on biased data, which it can then replicate and even amplify. While human impartiality cannot be guaranteed either, no algorithm can replace human

empathy, discretion, or judgement – a cry for pain, a plea for medical help, or a request to contact loved ones must never be dismissed or filtered out by an algorithm as background noise. Impartiality does not take place in a vacuum; it depends on systems deliberately designed to uphold it. Access to healthcare and family contact are rights that must be ensured, and the obligation to respect these rights is on humans, not on machines.

Detention, at its core, is about people. Maintaining good order and discipline requires meaningful and regular interaction between detainees and the guard force to sustain situational awareness and trust. Outsourcing the management of a place of detention to an algorithm risks removing our collective humanity from the equation.

Humans must therefore remain part of the feedback loop between inputs, processing, outputs and decisions, and keep the positive obligation of humane treatment at the centre of any calculation. This includes training guard forces not only on their legal obligations and detainee management techniques, but also on how to integrate AI effectively and responsibly into their work.

To ensure a truly human-centred approach, military planners, detaining authorities and AI developers must start with a shared understanding of the detention environment and its inherent vulnerabilities. Challenges related to humane treatment, conditions of detention, access to basic services, and personal data protection must be addressed from the design phase of any AI system onward.

Planning for compliance

In [a recent post](#), Isabelle Gallino and Sylvain Vité outlined the preparatory measures required for states to comply with IHL in detention operations during international armed conflicts. They highlighted how infrastructure, institutions, and instructions must be adapted well before hostilities begin. As AI enters the detention sphere, the same principles apply. To deploy IHL-compliant AI technologies, states must act in peacetime – designing detention/internment facilities and AI systems with compliance in mind, establishing clear legal and policy frameworks, and defining accountability among public institutions, private actors, and individuals.^[3]

Preparing for IHL compliance also means rigorously testing and adapting any AI technology under consideration. This is particularly important when states repurpose systems developed for other contexts. With criminal justice systems worldwide experimenting with AI tools, states may be tempted to import them into armed conflict detention operations, seeking efficiency. Yet as Ashley Deeks cautioned in [her post](#), doing so risks a “portability trap” – the failure to grasp how algorithmic solutions built for one social (or legal) environment may distort outcomes in another.^[4]

No AI system enters a detention setting as a blank slate. It interacts with existing laws, infrastructure, and social dynamics, from facility design to prisoner-guard relations. An algorithm trained on data from one population in a different cultural, social or legal context may yield biased or misleading results when applied to another. A second, legal portability trap can arise when AI developed under a human-rights framework or for high-security prisons is transferred to armed-conflict detention, where distinct rules and principles apply, including the assimilation principle^[5] for prisoners of war.

Training on and with AI systems is important both to test such technologies and to mitigate some of the risks described above. This may be done in military exercises that realistically replicate conditions of large-scale and high-intensity armed conflict. But considering the limitations inherent to testing in the controlled environment that is a military exercise,^[6] such testing may not, alone, constitute an effective safeguard. In fact, a military exercise will never account for all possible scenarios, for the adversarial conditions of armed conflict.

In addition to military exercises, states may “shadow test” technologies in real-time-real-life operations, precluding the AI system from having any decision-making or even decision-assisting capabilities. Such testing may allow relevant experts

to review and assess the outputs of the AI decision-support system (AI-DSS) post facto, without any risk of engaging personal or state responsibility and, most importantly, without any risk of violating the rights of the people on whom it is being tested. Based on the assessment of such testing, the technology and its use may be adjusted, adopted, or discarded.

Setting the necessary preconditions for compliance with the non-discrimination provisions of IHL^[7] requires a bias-aware approach from the outset – mitigating, reducing, and addressing bias at every stage of an AI system’s lifespan, from conception and training to deployment and review. Only by planning and preparing for compliance in peacetime can states hope to respect the full scope of their obligations if war does come.

Legal reviews, transparency and the right to redress

Article 1 common to the four Geneva Conventions (CA1) requires all Parties to “respect and ensure respect” for the Conventions. It is impossible to ensure that the application of a new technology such as AI to an armed conflict detention operation respects the Conventions’ obligations without conducting a legal review.^[8]

To be effective, such reviews should not only be conducted at the inception and design stage, and prior to any initial roll-out of a new AI technology, but also whenever an existing tool is to be used for a different purpose, in a different operating context, or under a different legal framework (think international vs. non-international armed conflict (IAC vs. NIAC), or the applicability of international or regional human rights frameworks). Some IHL rules on detention differ significantly between IAC and NIAC, and standards of treatment need to be adapted to the person’s personal situation, including gender, age, and disability. Threat assessment algorithms developed with data from a particular population cannot be assumed to be transposable to another without careful adaptation. In other words, legal reviews cannot be “one and done”; they must be continuous or repeated.

In the case of AI-DSS applied to detention operations in armed conflict, the legal obligations against which compliance must be reviewed will notably include humane treatment, non-arbitrariness of detention, non-discrimination, the effective implementation of judicial safeguards, as well as the relevant detention regime (internment under the Third or under the Fourth Geneva Convention, criminal justice, etc.).

Taking the example of AI-DSS applied to the decision to intern persons protected under the Fourth Geneva Convention (GC IV), it would not only be necessary to ensure that the use of such technology is compatible with the requirements of “absolute necessity” for the “security of the Detaining Power” (for internment on a belligerent’s own territory) and of “necessary, for imperative reasons of security” (for internment on occupied territory),^[9] but also that the procedural safeguards^[10] can be applied effectively, that the decision is non-discriminatory, and most importantly, that the determination is *individual*.

Reviewing the legality of AI systems against the procedural safeguards outlined in Articles 43 and 78(2) of GC IV would notably require ensuring that the use of AI-DSS does not run counter to the internee’s right to challenge the internment decision; including, for instance, access to information about the reasons of internment.^[11] The opacity built into AI-assisted decision making (often referred to as the “black box effect”) inherently limits the ability of the user – in this case the detaining authority – to fully grasp the process having led to a specific recommendation. In the context of armed conflict internment, we must also consider the authorities’ decision to withhold relevant information from the person being interned, for reasons of military necessity and confidentiality.

In a sense, this conjunction of factors may create *double* black-box effect for the person being interned. As *Pejic notes*, access to information about the reasons for their deprivation of liberty may “constitute an element of the obligation of humane treatment, as a person’s uncertainty about the reasons for his or her detention is known in practice to constitute a source of

acute psychological stress.”^[12] As such, the use of AI systems that are unable to provide internees with meaningful information about the reasons for their detention is unlikely to be IHL compliant.

AI-supported systems in armed conflict detention settings may be used to manage the flow and movement of people deprived of liberty within a place of detention. Such a system may allow to reduce the number of guards needed and to reallocate human resources to other tasks (including to combat functions). However, if applied to GCIII internment of prisoners of war, any system would need to allow its user to implement the internment logic, a logic of non-punitive deprivation of liberty, not in close confinement, with extensive freedom of movement within the perimeter of a camp. It goes without saying that an identical system applied to a non-international conflict detention reality would not be assessed and reviewed against the same standards, and that those systems would in no way be interchangeable without significant adaptations.

Conclusion

International humanitarian law is not the enemy of innovation or progress. On the contrary, it provides the guardrails and safeguards that enable developers and users to apply to mitigate risk and enable the responsible use of AI in the military domain. As states look to AI to leverage data, achieve efficiency gains and maximize resources in military operations, this cannot be done at the expense of the rights and dignity of persons deprived of liberty – or of their existing legal obligations.

With over 130 armed conflicts worldwide, and with corrosive interpretations of IHL sometimes used to justify non-compliance and explain away immense humanitarian consequences, we cannot assume that AI will always be deployed in detention operations with the best of intentions. And even if it is deployed with the best of intentions, it can still fall foul of the law. Like any other tool, it is how humans conceive, develop, test, and use it that will determine whether it becomes a force for good, or a vector for further violations and further suffering.

References

[1] Common articles 1, and 3 to the four Geneva Conventions of 1949

[2] ‘ICRC Position Paper: Artificial intelligence and machine-learning in armed conflict: a human-centred approach’, International Review of the Red Cross No 913, March 2022: “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.” <https://international-review.icrc.org/articles/ai-and-machine-learning-in-armed-conflict-a-human-centred-approach-913> .

[3] French Red Cross, Australian Red Cross, ICRC, *Private Businesses and armed conflict, and Introduction to relevant rules of international humanitarian law*, 4829/002, 2024, section 2.2.

[4] Andrew D. Selbst et al., ‘*Fairness and Abstraction in Sociotechnical Systems*’ In FAT* ’19: Conference on Fairness, Accountability, and Transparency (FAT* ’19), January 29–31, 2019, p. 61.

[5] The principle of assimilation is the foundation on which a number of rules in the Third Geneva Convention rest. It reflects an understanding that prisoners of war will be treated on the same terms as members of the armed forces of the

Detaining Power. See for Commentary of GCIII, ICRC, 2020, paras. 30–38.

[6] Robin Geiss, Henning Lahmann, ‘The use of AI in military contexts: opportunities and regulatory challenges’, *The Military Law and the Law of War Review*, Vol. 59 No. 2, 2021, pp. 182–3.

[7] CA3; GCIII, Art. 16.; GCIV, Art. 27.

[8] Note that Protocol Additional 1 to the Geneva Conventions (AP1) sets out a separate obligation to conduct legal reviews of any new means or method of warfare (Art. 36 AP1). States may therefore use similar processes and structures to review the legality of AI technology, including when used in detention operations.

[9] Article 42 and 78 GCIV, respectively.

[10] outlined in Articles 43 and 78(2)

[11] As outlined in Article 75(3) of Additional Protocol I; See also Jelena Pejic, ‘Procedural principles and safeguards for internment/ administrative detention in armed conflict and other situations of violence’, *International Review of the Red Cross*, Volume 87 Number 858, June 2005, p. 384: “The information given must also be sufficiently detailed for the detained person to take immediate steps to challenge, and request a decision, on the lawfulness of the internment/administrative detention”

[12] Pejic, p. 384.

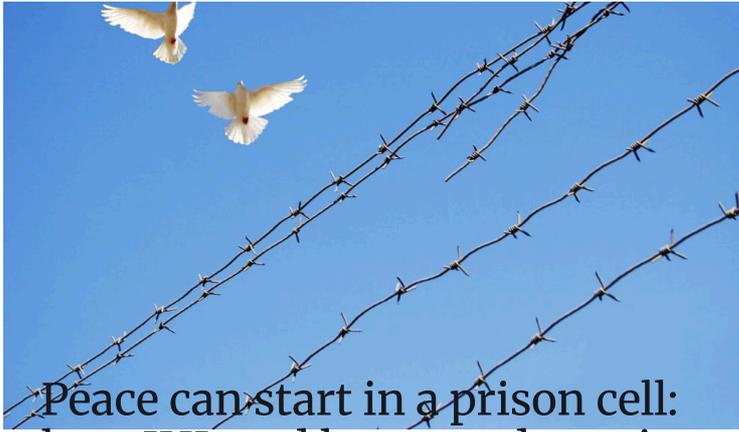
See also:

- Laura Bruun and Marta Bo, *‘Constant care’ must be taken to address bias in military AI*, August 28, 2025
- Wen Zhou and Anna Rosalie Greipl, *Artificial intelligence in military decision-making: supporting humans, not replacing them*, August 29, 2024
- Ruben Stewart and Georgia Hinds, *Algorithms of war: The use of artificial intelligence in decision making in armed conflict*, October 24, 2023

Tags: AI, AI decision support systems (AI-DSS), Artificial Intelligence, compliance, decisipo, Detention / Detainees, GCIII, GCIV, Geneva Conventions, human rights, humane detention, humane treatment, ill-treatment, international humanitarian law, non-arbitrariness of detention, Prisoners of War, Respect for IHL, Techplomacy

You should also read these articles





Peace can start in a prison cell: how IHL and humane detention can build pathways to peace

🕒 14 mins read

Accountability / Analysis / Detention / Generating Respect
for IHL / IHL / Law and Conflict / New Technologies

Terry Hackett & Audrey Purcell-O'Dwyer

When wars end, peace rarely begins overnight. It's built, slowly
and painstakingly, through acts that ...



From hackers to tech companies: IHL and the involvement of civilians in ICT activities in armed conflict

🕒 10 mins read

Accountability / Analysis / Detention / Generating Respect
for IHL / IHL / Law and Conflict / New Technologies

Tilman Rodenhäuser, Samit D'Cunha, Laurent Gisel, Anna
Rosalie Greipl & Marco Roscini

Picture a potential future armed conflict: missiles and drones
crowding the skies, uncrewed vehicles rolling ...