AI SAFETY AND INTERNATIONAL LAW

A Geneva Convention for AI and Robotics: Human Control, Civilian Protection, and Remedy

Eight proposed articles for autonomous systems in armed conflict and high-impact civilian services

Revanth Matha · Matha Labs | Preprint awaiting review 1.0 | 2026-11-20

Abstract

AI and robotic systems can affect people during armed conflict and through ordinary civilian services. Their lawful use requires rules about authority, protection and redress that apply to the institutions responsible for them. We examine primary treaty texts, humanitarian recommendations and recent diplomatic documents, then propose a standalone Geneva convention with separate conflict and civilian provisions. Eight candidate articles address scope, categorical prohibitions, continuity of essential services, effective human control, inspection, incidents, remedies and ratification. In three constructed cases, the draft permits bounded hospital logistics, requires immediate assistance and review after a water kiosk excludes a resident, and prohibits autonomous anti-personnel use in armed conflict. These dispositions demonstrate the internal operation of the draft and identify evidence a reviewer would need. Its prospects for adoption, implementation costs and effects on human outcomes remain open. The working text is an independent preprint for public review.

Keywords: human control, international humanitarian law, civilian robotics, human rights, accountability, treaty design

Contributions

1. Eight model articles that define the distinct triggers for armed-conflict and civilian duties.

2. A procedure connecting meaningful human control to an independent inspector, service continuity and individual relief.

3. Three worked legal cases and a comparison with a narrower protocol under the Convention on Certain Conventional Weapons.

1. Plain-English summary

A hospital can use a robot to deliver a meal. A resident who never receives it still needs someone to answer the phone, find out what happened and bring food. Now imagine a relief agency using AI to decide whose request to put at the front of a queue during fighting. The question becomes more serious, but the equipment may look ordinary. A treaty should identify who owes a duty to the person, what they must do before handing over a task, and what happens when the task goes wrong.

I propose a standalone Geneva convention with two parts. One sets prohibitions and responsibilities for AI and robotics in connection with armed conflict. The other governs civilian systems that exercise physical control over people or decide access to a service necessary for life or safety. A small hospital could keep its delivery cart. It would need a safe fallback if the cart holds up essential care. A state could not delegate a decision to direct force against a person to an autonomous weapon under this proposal.

The draft gives an injured person someone to approach without having to identify the model or prove how its internal calculations worked. An independent reviewer can inspect the evidence and order relief. The proposal also imposes duties on governments that procure or license systems. Its feasibility depends on whether a government can fund those duties and whether officials will obey an order to stop. No field trial, diplomatic agreement or estimate of injuries prevented is reported here.

2. What the documents can establish

This is a legal and policy design study. I selected primary texts that answer four questions. Which protections already apply in armed conflict? Which human rights and accessibility safeguards matter to a civilian operator? Which institutions have proposed additional rules? How could a new instrument bind a state? The source set comprises treaty texts, an institutional position, a General Assembly resolution, an account from the United Nations Office at Geneva and a dated working paper submitted by states. This is a purposive selection. It cannot establish consensus across governments or completeness across domestic legal systems. [1][2][3][4][5][6][7][8][9][10][11]

For each source I examined its legal character, the actor it addresses, the setting that triggers it and the decision it leaves open. A provision in an existing treaty is interpreted subject to its own scope and participation. An institutional recommendation records the position of its author. I used those distinctions to draft eight candidate articles. I then applied the same articles to three deliberately different, fictional implementations. The analytic result is the disposition in each case and the evidence that could change it.

The diplomatic record used here stops with the working paper dated November 18, 2026. The UN reports that on September 4 the expert group agreed on elements for a possible instrument while leaving its legal form open. Seven governments subsequently submitted a proposal to negotiate a protocol at the November 16-20 review conference. Proposed decisions in that paper remain proposals in this analysis. The General Assembly resolution and the ICRC position have different institutional roles. None of these texts adopts the convention printed here. [3][4][5][6]

Table 1. Legal authority and the decisions still open

Source	What it supplies	Question for a new convention
Geneva Conventions and Protocol I	Conflict protections and, for Protocol I parties, new weapons review [1][2]	Who verifies an updated system or provides an accessible complaint?
ICCPR and disability-rights convention	Applicable rights, remedies and accessibility obligations [8][9]	How does someone get immediate human reconsideration?
Council of Europe framework	AI lifecycle obligations and a national defence exclusion [7]	Which high-impact physical functions require a safe fallback?
ICRC, UN and state papers	Recommendations, deliberation and proposed negotiations [3][4][5][6]	What new wording could states adopt and implement?

Each source is used for the legal question it can answer. The third column identifies a drafting choice in this preprint.

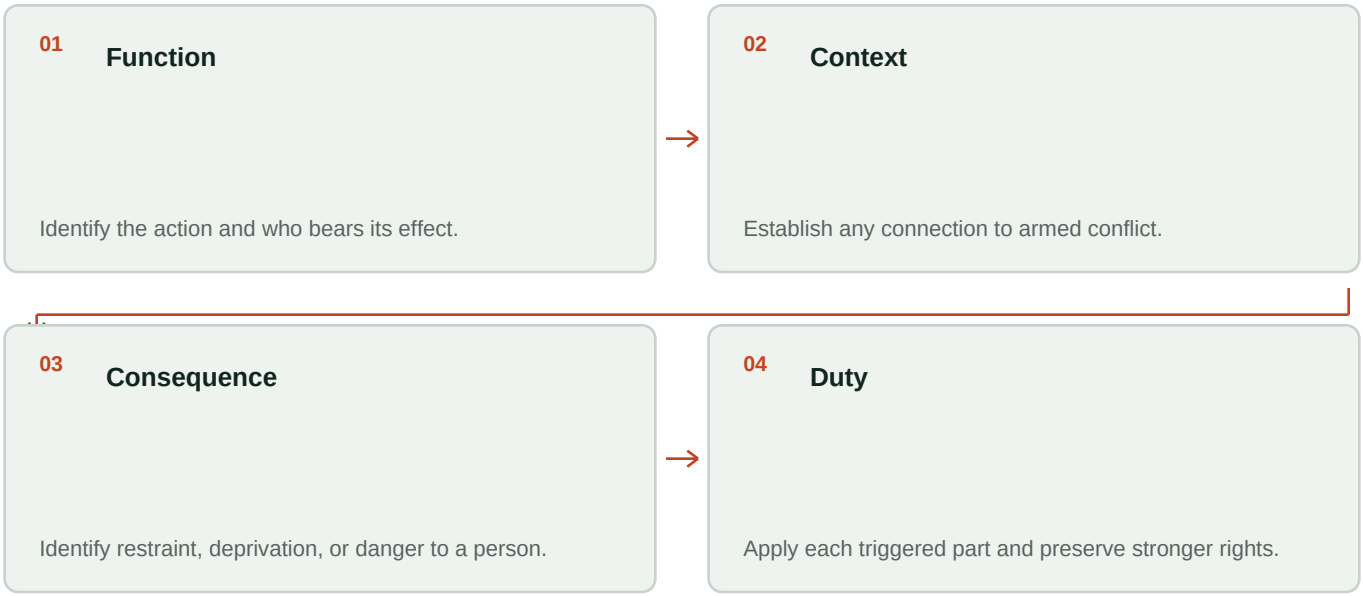
3. Article 1. Scope, definitions and applicable law

Proposed operative language. Each Party shall give effect to Part I for covered uses connected with international or non-international armed conflict. It shall give effect to Part II for covered civilian functions within its jurisdiction, including public functions contracted to private entities. Both parts may apply to one operation. A Party shall preserve stronger protections under applicable international and domestic law. A person working for the state, a municipality or a contractor may receive a duty under domestic implementing legislation.

For the proposal, an AI system infers an output from its inputs. A robot physically acts in its environment. Covered functions are defined by what happens to people. The conflict part covers the delegation of force and decisions materially affecting the treatment of protected persons. The civilian part covers restraint or confinement, or a consequential determination that can deny, delay or terminate access to an essential service. An essential service is one whose interruption in the circumstances presents a material danger to life or bodily integrity. Manufacture, placement and supervision of covered equipment trigger the relevant preventive duties.

The same machine can have several functions. A warehouse robot recording stock movements generally falls outside these special civilian provisions. Connecting that record directly to a rule that locks residents out of their only drinking-water supply brings the decision and the lock within them. Ordinary workplace and product-safety duties continue to matter in both settings. This distinction permits a reviewer to examine a concrete function, including one implemented by simple software, without debating the intelligence of the machine. Where parts overlap, the civilian remedy remains available and cannot license an act prohibited under the conflict part.

Figure 1. Deciding which duties attach



Read the four questions in order. Setting and consequence are separate inquiries. A civilian hospital in an area of armed conflict can engage both parts for the same decision.

4. Article 2. Prohibitions and protection during armed conflict

Proposed operative language. Each Party shall prohibit the use of autonomous weapon systems designed or used to apply force against persons. It shall prohibit the use of autonomous weapon systems whose effects cannot be sufficiently understood, anticipated and explained for a responsible authority to establish compliance with applicable law in their intended circumstances. A favorable average test result shall provide no exception to these prohibitions. Parties shall prevent their agencies and contractors from authorizing prohibited uses and provide for investigation of credible allegations.

These proposed prohibitions draw on the ICRC's recommendations. The 1949 conventions already protect people under defined conditions. Common Article 3 sets minimum guarantees in a non-international armed conflict, including humane treatment for people taking no active part in hostilities and care for the wounded and sick. Protocol I, in its scope, addresses distinction, protection of civilians and precautions. Its Article 36 requires a party considering a new weapon, means or method of warfare to examine whether its employment would be prohibited in some or all circumstances. A state must identify the legal rules applicable to it in the particular situation. The proposed text adds specific prohibitions. It offers no technical instructions for the use of weapons. [1][2][3]

People already in the care or power of an authority need a separate protection. Under this draft, an AI assessment could alert a medical team to an unmet need. It could never supply the sole effective decision to withhold humane treatment, urgent medical review or essential humanitarian assistance from a protected person. A designated human official must be able to reverse or disregard an output while relief can still make a

difference. For non-state armed groups, existing applicable humanitarian obligations retain their own legal basis. Ratification by one government does not itself make every armed group a party to a new convention. [1][11]

5. Article 3. Civilian use and continuity of essential services

Proposed operative language. Each Party shall require the person or organization exercising a covered civilian function to offer an accessible route to a human decision, maintain continuity of urgent services and provide reasons appropriate to the affected person. Before putting a system into service, the responsible body shall demonstrate how someone can reach the human alternative if the system, its communications or its usual staffing arrangements fail. The obligation applies to a private operator entrusted with the relevant public function.

A robot helping an older person move safely may act promptly on instructions given by that person. Forcing the person to wait for an official to approve every movement would itself undermine autonomy. The draft instead protects the user's ability to direct assistance, decline it and obtain accessible help. A facility deciding to confine that person or refuse requested essential care faces a different rule. A high prediction score about fall risk cannot itself create legal authority for confinement. Any restriction needs independently lawful grounds, an accountable decision and timely reconsideration. Articles on accessibility, consultation and access to justice in the disability-rights convention inform that distinction. [9]

The Council of Europe Framework Convention offers a useful comparison. It addresses AI lifecycle activity, accessible remedies and risk assessment, and excludes matters relating to national defence in Article 3(4). The human-rights covenant supplies protections including life, privacy and an effective remedy within its applicable scope. The civilian part proposed here specifies a physical fallback and an immediate human alternative for a narrower class of decisions. It derives its obligations from express consent to the proposed treaty. Applying international humanitarian law to everyday peacetime robotics would supply the wrong legal basis. [7][8][9][11]

6. Article 4. A human with the authority and time to act

Proposed operative language. Each Party shall require a named public or civilian duty holder for each covered function. For each permitted deployment, the holder shall identify the decision reserved to a person, the information needed to make it, the period within which intervention matters and the resources for continuity if the system is suspended. A material change in the function, population, operating conditions or supplier shall require review before the revised use proceeds. No designation of a supervisor shall permit an act prohibited by Article 2.

Consider a city distributing drinking water during a prolonged outage. The duty holder can demonstrate human control by having the reviewer observe a resident seeking help in person, a staff member making a decision and another worker delivering water while a disputed restriction is examined. An employee with theoretical authority but a queue longer than the time available for relief cannot deliver this protection. The assessor should record the time from request to effective assistance and the cases that went unserved. Measuring how rapidly a screen displayed the request would answer a different question.

Responsibility must survive changes of shift and changes of supplier. The purchasing body identifies an alternate officer with the same powers and documents who may order a pause, who performs the manual service and who may authorize a restart. A worker who raises a credible safety concern receives protection against retaliation. The affected person may use an accessible channel without an account or a smartphone. Local authorities set deadlines according to the danger of the delay and publish the basis for those deadlines. An urgent request receives immediate human triage rather than an ordinary administrative waiting period. These are specific proposed duties inspired by wider remedy obligations. [7][8][9]

7. Article 5. Inspection, confidentiality and participation

Proposed operative language. Each Party shall designate or establish a functionally independent authority empowered to inspect covered operators, obtain necessary records, hear staff and affected people confidentially, issue binding preservation or suspension orders and require evidence before resumption. Parties shall fund this authority independently of any supplier it oversees. Parties shall provide a review procedure for such orders that preserves urgent safeguards while protecting due process.

The inspector starts with questions a person can answer. Who commissioned the operation? Who changed what the system was allowed to do? Could an operator provide urgent assistance when a complaint arrived? What evidence could contradict the operator's account? A written policy establishes that someone designed a procedure. A recorded, supervised continuity exercise can establish that staff implemented it under the conditions exercised. A procedure may fail at night, at a remote facility or for a resident who needs an interpreter. Inspectors should sample those conditions in consultation with affected communities. A one-time certification cannot establish continuing performance after a material change. [7][9]

Affected people and operators both have interests in how evidence is handled. Storing every conversation or identifying every aid recipient would create privacy exposure and could deter people from seeking assistance. The duty holder should retain the authorization, system version, material decision, available reasons, complaint disposition and time to remedy for a defined period. Preserve additional material under a lawful specific order. Where disclosure could endanger people or compromise a legitimate protected interest, an independent reviewer can examine protected material and provide a sufficiently informative summary. A government cannot defeat oversight merely by marking an entire operation secret. A complainant should be able to challenge a closure and a provider should be able to contest a mistaken finding. [7][8]

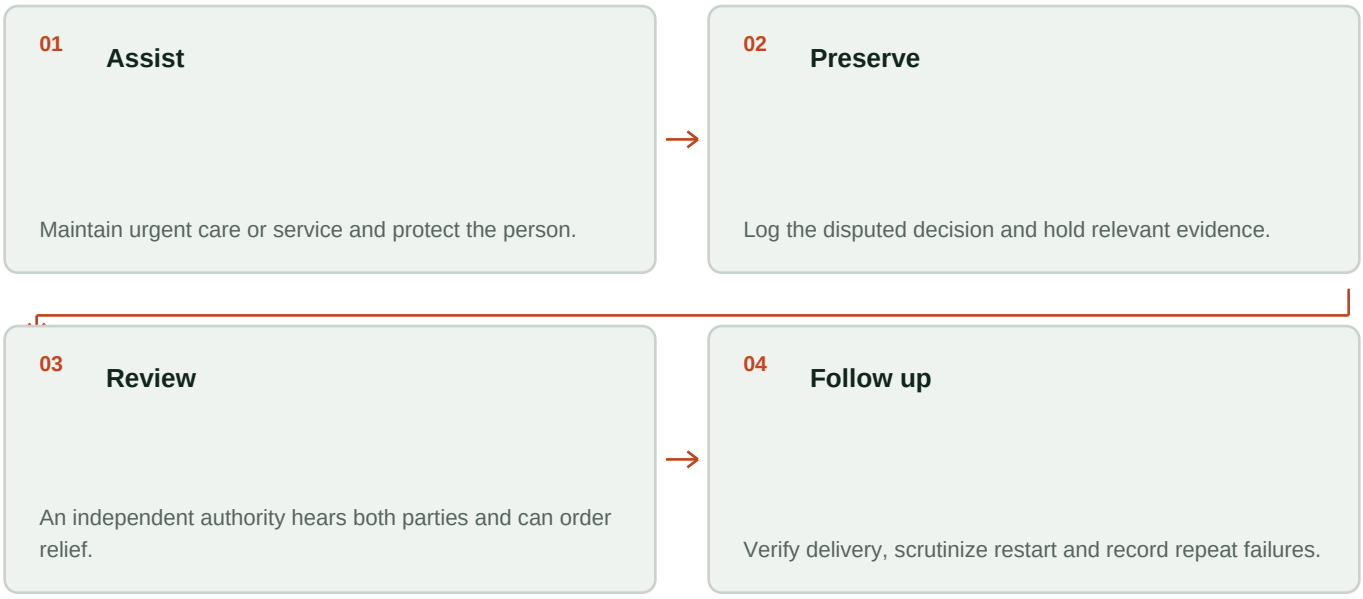
8. Article 6. Incident reporting and continuity during an investigation

Proposed operative language. Each Party shall require prompt notice of a serious incident or credible near miss to the independent authority. Operators shall provide immediate assistance, preserve designated evidence and suspend the affected function where continued operation creates a material risk of serious harm. Restart requires a recorded decision by a competent authority empowered to reject the proposed correction. Public reporting shall protect personal information, humanitarian confidentiality and witnesses.

Suspending a robot should not suspend a patient's access to food, shelter or treatment. The authority can order an unsafe automated decision process stopped while a human team continues the service. Decisions about stopping, substitution and restarting have different owners and different evidence needs. An emergency plan that says only to switch the machine off can itself endanger people dependent on the service. The proposal requires the facility to provide and test the substitute in advance.

A conference of the parties would receive anonymized national reports. The meaningful denominator might be the number of affected service decisions, the number of people dependent on a single site or the number of occasions on which a fallback was needed. Any rate should display its denominator, period and coverage. A low complaint rate can arise because people fear punishment or because the filing process excludes them. The monitoring body would seek feedback through disability organizations, independent local institutions and humanitarian representatives with appropriate confidentiality. This mechanism reports evidence on implementation. A reduction in harm would require a separate assessment with a credible comparison. [7][9]

Figure 2. What happens after a person reports harm



Immediate aid can begin before attribution is settled. The later inquiry examines responsibility, verifies compliance with orders and protects the opportunity to appeal.

9. Article 7. Responsibility, remedy and the person who bears the loss

Proposed operative language. Each Party shall provide accessible procedures for interim relief and a reasoned independent determination. Available remedies shall include cessation, restoration of service, correction of records and compensation where warranted. The state shall ensure a route to relief where a covered public function is contracted out or its supplier becomes insolvent. Parties shall establish appropriate administrative, civil and, for conduct defined by law, criminal consequences. No model or machine may discharge a duty assigned by this convention to a state, organization or person.

A person denied urgent water should be able to identify the service and the harm, request immediate assistance and have an independent reviewer secure records held by the agency or contractor. The claimant may have no access to the vendor and no way to reproduce a system output. An institution with exclusive access to a relevant record should have to explain its absence. A court can decide the proper civil consequence of missing evidence, with both sides heard. That proposal for civil procedure does not reduce the presumption of innocence in criminal proceedings. [7][8]

Public responsibility for procuring a system, civil liability for harm and individual criminal liability answer different questions. The Rome Statute assigns individual liability within the court's jurisdiction to natural persons and lays down specific legal requirements for responsibility of superiors and mental elements. Signing a procurement order or holding a supervisory title alone cannot establish a crime under that statute. A new convention cannot enlarge the court's jurisdiction by announcing that it has done so. A domestic offence would require clear legislation, prospective application and a fair proceeding. Administrative sanctions and repair of the harm remain available under their own standards even where criminal liability cannot be established. [8][10]

10. Article 8. Ratification, assistance and compliance across borders

Proposed operative language. The convention shall open for signature by states and permit accession. It shall enter into force ninety days after deposit of the twentieth instrument of ratification, acceptance, approval or accession. For each state joining thereafter, it shall enter into force ninety days after its deposit. No reservation shall be permitted to the prohibitions in Article 2 or the minimum right to urgent assistance and independent review in Article 7. These thresholds and timelines are proposed bargaining terms. A conference of the parties shall arrange implementation assistance and examine compliance reports.

The title expresses an ambition for a diplomatic conference in Geneva. Adoption of this text would require negotiation, consent and entry into force under its agreed terms. The draft would stand alongside the four 1949 conventions. It would preserve their protections and leave their existing text intact. A state that joins implements duties concerning covered public agencies, suppliers and operators through domestic law. A private supplier does not become a party by delivering software to a government, and a non-party state does not acquire treaty duties simply because a neighboring state ratifies. [1][11]

Cross-border contracting poses a practical test. The client state can condition its own procurement on enforceable cooperation with an inspector, access to necessary documentation and continuity if a supplier exits. Participating states can recognize preservation requests and cooperate through specified national contact points. Overseas inspection would require a treaty basis, agreement or another lawful channel. The proposal gives peer reviewers access to agreed evidence and the authority to publish reasoned findings after hearing the responding state. It supplies no unilateral right of entry. Lower-income parties would need pooled translation services, shared technical assessors and funding for manual continuity. Reporting duties with no implementation budget could divert scarce health and relief resources from the people the convention seeks to protect.

11. Three decisions under the draft

The first constructed case concerns a medical supply cart in a hospital treating people during armed conflict. The cart transports sealed supplies between rooms. Nurses can retrieve the supplies and prioritize individual patients. Under those facts, transportation may proceed subject to ordinary protections and the relevant continuity review. If a software update makes the cart the sole means of receiving an urgent item and staff can no longer override its refusal, the supply process now decides access to care. The facility must restore a human route immediately. The same hardware yields a different assessment because its assigned authority changed.

The second case concerns an automated municipal water kiosk in peacetime. Its identity service rejects a resident who has no current document during a prolonged outage. The proposed civilian provisions require urgent triage and a way to obtain water while the disputed decision is reviewed. An operator may keep an ordinary records check for later reconciliation. The independent reviewer can examine the restriction, direct repair of an erroneous record and scrutinize effects on people without accessible documentation. The conflict prohibitions do not supply the legal trigger here. The resident's dependence on this source and the consequences of interruption do.

The third case concerns an authority proposing an autonomous anti-personnel weapon in an armed conflict. The first proposed prohibition forecloses the use. A favorable safety rating, emergency staffing shortage or supervisory sign-off cannot authorize it. The competent authority records the proposed use, prevents its authorization and refers any suspected violation to an independent investigation with protection for witnesses. This example illustrates a legal boundary. It makes no empirical claim about any particular system. [3]

All three results are deductions from the draft articles and the stated facts. The cases were chosen to expose whether a benign function remains available, whether a civil right can be protected promptly and whether a categorical prohibition remains categorical under pressure. They provide a way for reviewers to find a contradiction. They cannot show how frequently such cases occur, whether the institutions actually exist or whether states would comply.

Table 2. Results of three constructed cases

Case	Proposed disposition	Evidence needed
Hospital supply cart	Permit transport with a workable urgent-care fallback	Staff demonstration, service access and change history
Water kiosk blocks resident	Provide interim water and independent reconsideration	Urgency, human route, reason for refusal and restoration
Proposed autonomous anti-personnel use	Prohibit authorization and investigate credible breaches	Documented purpose, authority's decision and independent review

The examples evaluate internal consistency of the proposed articles. They are hypothetical. No patient, resident, military unit or field operator took part.

12. Competing treaty designs and conditions for revising the proposal

A new protocol under the Convention on Certain Conventional Weapons could build on existing diplomatic work and concentrate expertise on autonomous weapons. Seven states have proposed that course. It would be a natural setting for a narrower prohibition, and this paper would support progress there even if the wider convention never attracts enough states. The drawback for the present aim is institutional scope. A weapons forum cannot be assumed to regulate municipal water service, healthcare robotics and private operators outside armed conflict. A parallel civilian instrument could address those issues at the cost of two sets of negotiations and a more difficult boundary between them. [5][6]

A single mixed convention makes common duties, including independent review and continuity, visible in one place. It can also slow adoption. A state prepared to prohibit a narrowly defined weapon might reject a treaty that changes how cities provide relief. A lower-resourced signatory might accept principles it cannot enforce because it lacks trained inspectors, interpreters or backup staff. A peacetime rights advocate might object that a body dominated by state security officials cannot hear civilian grievances impartially. These are design objections with concrete implications. Civilian expertise on the review panel, independent funding, separate reporting by function and domestic access to courts would be necessary for the proposed structure.

Negotiators should compare the text with a narrower conflict protocol and a civilian implementation agreement during open consultations. A test panel should include people with disabilities, displaced people, humanitarian workers, small public operators and inspectors representing different legal systems. Ask each group to apply the text to the same scenarios, identify who can act during an interruption and identify the information a claimant can safely obtain. Record disagreements and unresolved questions. Revise the text if officials cannot distinguish routine logistics from a denial of care, if claimants must obtain vendor secrets to be heard or if inspection would expose people to retaliation.

This preprint does not estimate costs, test a treaty in operation or show that a single instrument would secure broader ratification. Its contribution is a draft with identifiable duties, limits and review routes that can be

challenged article by article. The most important practical question remains simple. When a machine makes a grave mistake affecting a person, which institution can give that person help soon enough, and what gives it the legal power to do so?

References

[1] Geneva Convention IV (1949), articles 2 and 3. Authentic treaty text on applicability and humane treatment, including Common Article 3. Accessed 2026-11-20.

[2] Additional Protocol I to the Geneva Conventions (1977), articles 36, 48, 51 and 57. Treaty rules concerning new weapons review and protection of civilians in international armed conflicts. Accessed 2026-11-20.

[3] International Committee of the Red Cross. Position on autonomous weapon systems (2021). Institutional recommendation for new prohibitions and restrictions, including unpredictable systems and anti-personnel use. Accessed 2026-11-20.

[4] United Nations General Assembly. Lethal autonomous weapons systems, A/RES/80/57 (2025). General Assembly resolution recorded in the UN Digital Library. Distinct from a treaty or ratification. Accessed 2026-11-20.

[5] United Nations Office at Geneva. Seventh Review Conference of the Convention on Certain Conventional Weapons (2026). UN account of the September 4 expert report and the agenda of the November 16-20 review conference. Accessed 2026-11-20.

[6] Austria, Belgium, Brazil, Ireland, New Zealand, Panama and Mexico. CCW/CONF.VII/WP.7 (18 November 2026). Delegations' working paper proposing negotiations toward a legally binding protocol. Its annex contains proposed decisions. Accessed 2026-11-20.

[7] Council of Europe. Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, CETS 225 (2024). Treaty text, especially articles 3, 14-16 and 23 on scope, remedies, risk management and follow-up. Accessed 2026-11-20.

[8] United Nations. International Covenant on Civil and Political Rights (1966). Treaty text, especially articles 2, 4, 6 and 17 on remedy, derogation, life and privacy. Accessed 2026-11-20.

[9] United Nations. Convention on the Rights of Persons with Disabilities (2006). Treaty provisions on consultation, accessibility, situations of risk and access to justice. Accessed 2026-11-20.

[10] International Criminal Court. Rome Statute, articles 22, 25, 28 and 30. Consolidated text on legality, natural persons, responsibility of superiors and mental elements. Accessed 2026-11-20.

[11] United Nations. Vienna Convention on the Law of Treaties (1969), articles 11-16, 24, 26 and 34. Treaty framework for consent to be bound, entry into force, observance and the position of third states. Accessed 2026-11-20.

Reproducibility record

Private review materials: https://github.com/RevanthM/geneva-ai-robotics | Research snapshot: ce4cb1f9d94b66da03e4b5ae47ad3e5f7cab64e9