

MEcoVEA e.V.

ANALYTICAL REPORT / 2026

HUMAN CONTROL BEFORE LETHAL AUTONOMY

Cybersecurity, Accountability and
Technical Safeguards for AI-Enabled
Military Decision-Making

BEFORE WE TEACH AI
TO KILL, WE MUST
MAKE SURE IT CAN
STOP.

MEcoVEA e.V.

Berlin · 2026

Civil-society analytical report

THE SPEAKER

JURI BRUNNMEIER



Juri Brunnmeier is a cybersecurity and systems engineering specialist with experience in the cybersecurity and safety assurance of complex technical systems. He has served as Project Security Manager at Thales Deutschland and HITACHI Rail GTS Deutschland, focusing on cybersecurity, RAMS, and systems validation.

His background includes scientific research in semiconductor physics and quantum structures, with work presented within the Deutsche Physikalische Gesellschaft (DPG) and published in Physical Review B.

He is the founder and Managing Director of Brunotec GmbH, specialising in cybersecurity for industrial and critical infrastructure, IT/OT systems, technological security, and risk management, and heads the non-profit association Gnosis e.V.

For this report, Brunnmeier served as principal expert and lead contributor, focusing on military AI architecture, cybersecurity, system safety, human control, and technological risk management. He will present the report's findings on behalf of MEcoVEA e.V.

SPEAKER'S FRAMING

The question is not whether machines can process more data. The question is whether institutions retain the technical and legal capacity to stop an insufficiently grounded use of force.

CONTENTS

Executive Summary	5
Key Findings	6
Methodology and Limitations	7
Definitions and Analytical Scope	8
1. Why the Problem Is No Longer Theoretical	9
2. Digital Targeting Infrastructure: Palantir as an Illustrative Case	10
3. Speed, Uncertainty and the Boundary of Human Decision	12
4. Engineering Restraint under Uncertainty	13
5. A Theoretical Architecture for Safe Governance of Military AI	14
6. Cybersecurity, Observability and Independent Oversight	17
7. Assured Interruption of the Force-Application Loop	19
8. Accountability, International Humanitarian Law and the OSCE	21
9. Recommendations	25
10. Conclusion	29
Annex A. Minimum Assurance Checklist	31
Annex B. Terminology	32
Annex C. Proposed OSCE Confidence-Building Measures for Military AI	33
Sources	35
Notes	38

EXECUTIVE SUMMARY

A policy-facing summary of the technical, legal and institutional argument

Artificial intelligence in the military domain is moving closer to operational decision-making. AI-supported systems can fuse sensor data, classify objects, identify patterns, rank options and generate recommendations at a speed and scale that traditional staff processes cannot match. The decisive governance issue is therefore not limited to fully autonomous weapons. Algorithmic influence may become operationally significant earlier - when software determines what a human sees, which evidence is prioritised, how uncertainty is displayed and how much time remains for review. [1,2,3]

This report distinguishes analytical AI and decision-support systems from digital target-development workflows and from autonomous weapon systems. That distinction matters legally and technically. A system may leave formal authority with a human and still narrow the practical space for independent judgement. Human presence is not enough if the operator lacks time, source access, competence or a genuine ability to reject, delay or stop the recommended action. Human-factors research on automation bias and overreliance supports the need to treat interface design and operational tempo as part of the control problem. [31,32,33]

The report uses Palantir as an illustrative case of a broader technology class, not as an accused company and not as proof that a specific platform independently authorises lethal force. Public documentation shows target boards, high-priority target lists, intelligence-source fields and confidence indicators, as well as integrated AI and human-review workflows. These materials demonstrate how data, models and operational processes can be brought together; they do not establish the degree of automation or the role of any product in a specific strike. [22,23,24,25,26]

CENTRAL PROPOSITION

The safer design objective is not “make the model moral.” It is “make critical uncertainty technically consequential.” When mandatory conditions are missing or contradicted, the system should be able to refuse, escalate, degrade safely or block the effector channel.

The report therefore proposes an external safety envelope around high-risk AI functions. The model should not be the sole judge of its own correctness. A distinct control layer should verify formalised preconditions: permitted model and policy versions, user authorities, provenance and age of required data, mandatory independent confirmations, geographic and temporal restrictions, protected-object constraints and the approved order of authorisation. This does not automate international humanitarian law. Distinction, proportionality and feasible precautions still require contextual human legal and command judgement. [4,5]

Cybersecurity is integral to this architecture. A military AI system inherits ordinary software, identity, network and supply-chain risks while adding risks related to model integrity, data poisoning, adversarial input, sensor spoofing, prompt injection and model drift. Monitoring should therefore cover infrastructure, data, models and the decision process, with evidence preserved for forensic reconstruction. Secure development and risk-management frameworks such as NIST AI RMF and NIST SP 800-218A are useful reference points, although they are not military-specific certification regimes. [7,8,9]

A second core proposal is assured interruption of the force-application loop. The report deliberately avoids the fiction of a single universal red button. Interruption should be layered: advisory-only mode, blocking of the AI-to-effector interface, revocation of authorities or credentials, network isolation, deactivation of specific weapons functions and local emergency control. The safe state must be defined function by function, because disabling a protective function can itself create risk. DoD Directive 3000.09

similarly embeds governability, including the ability to disengage or deactivate systems that demonstrate unintended behaviour. ^[6]

The legal baseline remains human and State responsibility. Article 36 of Additional Protocol I requires States Parties to review new weapons, means and methods of warfare. Existing IHL is not displaced by AI, and the use of automated components cannot convert technical output into a legal determination by itself. At the multilateral level, the CCW process, UN General Assembly resolutions on autonomous weapons and AI in the military domain, the REAIM process, NATO's responsible-use principles and other national or coalition initiatives demonstrate a growing convergence around accountability, testing, governability and human control, even though important legal and political differences remain. ^[4,14,15,16,17,18,19,20,21]

For the OSCE, the practical opportunity is narrower and more realistic than technical certification. The Organisation can contribute through shared terminology, voluntary transparency, protected incident-exchange formats, contact points, discussion of national review procedures and confidence-building measures. Paragraph 24 of the OSCE Code of Conduct, requiring protection against accidental or unauthorised use of military means, provides a politically binding point of reference for this discussion; extending its logic to algorithmically unforeseen action is a proposal of this report, not an existing OSCE rule. ^[11,12,13]

KEY FINDINGS

01 Influence begins before autonomy

AI can shape the use of force through attention, ranking, classification and time compression even when a human still presses the final button.

02 Human-in-the-loop can be an illusion

Control is genuine only if the human has time, relevant information, competence and real authority to reject or stop.

03 Uncertainty must remain visible

A confidence score is not automatically a probability of correctness. Data uncertainty, model uncertainty and legal uncertainty should not be collapsed into one persuasive number.

04 The model must not police itself

High-risk systems need an external safety envelope with separated authorities, protected policies and the ability to block an inadmissible transition to the effector channel.

05 Disconnection must be layered

Safe interruption can mean advisory-only operation, interface blocking, credential revocation, network isolation or local deactivation - not necessarily a universal system shutdown.

06

Complexity raises the burden of accountability

If a State cannot review, constrain, stop and reconstruct the decision chain, technical opacity is not a justification for granting the system greater authority.

WORKING TEST

If the operator cannot explain why the system reached the recommendation, cannot see what would invalidate it, and cannot actually stop the process, “human control” is only formal.

METHODOLOGY AND LIMITATIONS

Scope

Desk-based analytical research on publicly documented military-AI decision support, autonomous weapon governance, cybersecurity, human factors and international humanitarian law.

Source hierarchy

Priority is given to primary and authoritative sources: ICRC, UN/UNODA, OSCE, Additional Protocol I, U.S. DoD, NIST, NATO and official vendor documentation. Peer-reviewed human-factors research is used for automation-bias and calibration issues.

Illustrative case

Palantir is used because its public documentation makes data-to-target workflows unusually visible. The case illustrates architectural possibilities; it does not establish actual operational configuration, combat effectiveness or autonomous authorisation of force.

Operational allegations

Where public reporting concerns concrete military operations, the report attributes the claim to the named source. The 2026 Iran example is based on a UN fact-finding mission and is not used to infer that a particular AI vendor caused the incident.

No classified access

The authors did not have access to classified systems, source code, rules of engagement, model weights, internal testing data or operational logs. Therefore the report does not evaluate the safety or legality of any specific deployed system.

Normative status

“Technical empathy” and “right to disconnect” are authorial concepts. They are used as shorthand for engineering restraint and assured interruption; they are not presented as established terms of international law.

Research cut-off

24 September 2026. The technology and policy environment changes quickly; future editions should re-check technical documentation, legal positions and multilateral processes.

DEFINITIONS AND ANALYTICAL SCOPE

The unit of analysis is the decision loop, not the model in isolation

This report treats military artificial intelligence as a family of software and socio-technical systems embedded in the preparation, taking and execution of military decisions. Risk depends less on the label “AI” than on the system’s function, operational setting and practical influence over the use of force.

Category	Operational function	Why the distinction matters
Analytical AI / AI-DSS	Fuses data, classifies objects, detects patterns or generates recommendations for a human.	May strongly frame attention and options without itself selecting and engaging a target.
Digital target-development / management	Supports identification, verification, prioritisation and approval of targets.	Software structures the path to a decision while formal authority can remain human.
Autonomous weapon function	After activation, selects and engages one or more targets without further human intervention.	Corresponds to the ICRC working definition of an autonomous weapon system; other autonomous functions do not automatically make a system an AWS. ^[3]

The key unit of analysis is therefore not an individual model but the entire decision loop: data sources, algorithms, processing rules, interfaces, user authorities, time constraints, authorisation procedures and the link to the effector system. Even if an algorithm merely sorts information or assigns priorities to objects, the risk is not limited to an ordinary analytical error. Excluding an object from the sample, misclassifying it, inflating a confidence score or visually ranking options can determine which information an operator sees first, what he or she regards as significant and which alternatives remain practically available. When several digital systems are combined, errors can accumulate and propagate through the whole planning and decision chain.

The transition from decision support to the de-facto delegation of a decision cannot be defined solely by the presence of a human who presses the final button. One must ask whether that person has sufficient time, information, professional competence and authority to independently verify the machine output, to reject it and to stop the process.

If those conditions are absent, a human may formally remain in the loop while in practice merely confirming a decision already shaped by the system. The ICRC explicitly warns of the risk of cognitive bias expressed as excessive trust in automated conclusions (automation bias) and of the reduction of the operator to a formal approver who simply rubber-stamps the system’s recommendation (human rubber stamp) when AI-DSS are used in armed conflict.

This report does not propose a ban on military artificial intelligence as such. That does not exclude the need for specific prohibitions or strict limitations on the most dangerous categories of autonomous weapon systems. Properly designed and employed AI-DSS can accelerate information processing, reduce the burden on personnel and, under certain conditions, help reduce harm to civilians. Yet such an

outcome must be demonstrated by testing, legal review and evaluation in the concrete environment of use - not assumed on the basis of the system's intended purpose or the claims of its developer.

ANALYTICAL THRESHOLD

the closer a machine output stands to the application of force, and the graver the possible consequences of error, the stricter the requirements must be for verifiability, traceability, operational constraints, human legal judgement, continuous oversight and the ability to intervene in time, to suspend the system or to place it in a pre-defined safe state. Operational effectiveness does not replace lawfulness, and the mere presence of a human at an interface does not prove that the decision remains genuinely human.

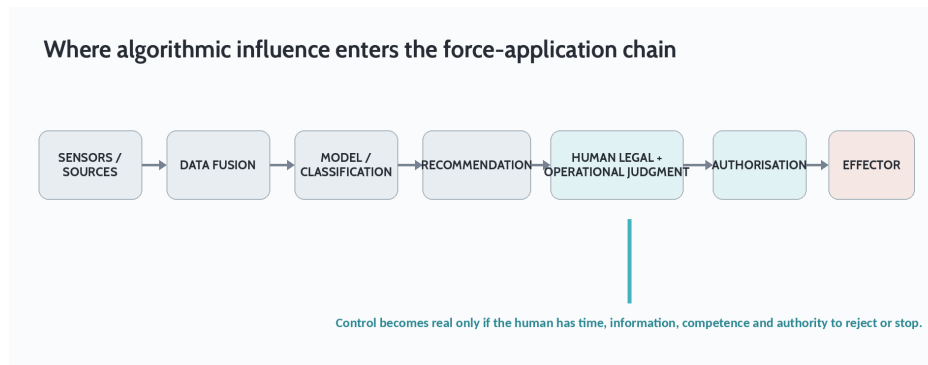


Figure 1. The closer an algorithmic output comes to the force-application channel, the higher the burden of verification, human judgement and interruptibility.

01 WHY THE PROBLEM IS NO LONGER THEORETICAL

Traditional staff procedures were never immune to error, bias or time pressure. Yet they at least left room, in principle, for comparing sources, professional argument, checking assumptions and documented coordination. A digital loop can compress these stages sharply: a sensor registers an object, a software layer links the observation to existing data, a model classifies the object, assigns it a priority and passes a recommendation to the operator. Such acceleration can increase tempo, but it can also reduce the time available for checking the original data, considering alternative hypotheses and assessing possible consequences.

Suppose an interface shows the operator a figure of “94 percent”. Without an explanation of how that figure was calculated, it can be read as near-certainty, even though in reality it may be no more than an internal model score. If the operator has only a few seconds, a high numerical value can replace substantive verification with a psychological sense of precision. Formally the decision remains human; in practice the algorithmic system has already framed the space in which that decision is taken. ^[30]

Particular danger arises from the combination of a visually persuasive machine output and the few seconds left to a human for confirmation. Legal authority may formally remain with the operator, yet the real possibility of exercising that authority becomes doubtful. Genuine human decision requires not only the presence of a person in the loop but also sufficient time, accessible information, professional competence, the possibility of independent verification and a real right to reject the recommendation or stop the process. When these conditions are missing, human-in-the-loop risks becoming human rubber stamp - a person who legally confirms a decision that has in fact already been shaped by the machine.

The ICRC explicitly links this risk to automation bias, which is especially pronounced in the stressful and rapidly changing environment of armed conflict. ^[31,32,33]

The international agenda confirms the urgency of the problem while demanding precision of description. In April 2025 the ICRC submitted a dedicated analysis of AI in the military domain, including AI-supported decision-making and the risks of automation bias and cascading error.

The ICRC has also clarified its legal position on autonomous weapon systems. In December 2024 the UN General Assembly adopted resolution 79/239 on AI in the military domain, opening a broader multilateral process beyond autonomous weapons alone. On 7-8 May 2026 the Swiss OSCE Chairpersonship held the conference “Anticipating technologies - for a safe and humane future” in Geneva. These developments demonstrate institutional relevance; they do not amount to a single agreed international policy on military AI. [1,3,16,12]

What is at stake, then, is no longer futurology but the gap between the speed of technological introduction and the capacity of law, military command and oversight mechanisms to ensure a genuinely human decision. The danger is not that a machine has already become a legal subject of decision-making; the danger is that a technical loop can turn human judgement into a procedural formality before institutions have learned to distinguish real control from its imitation.

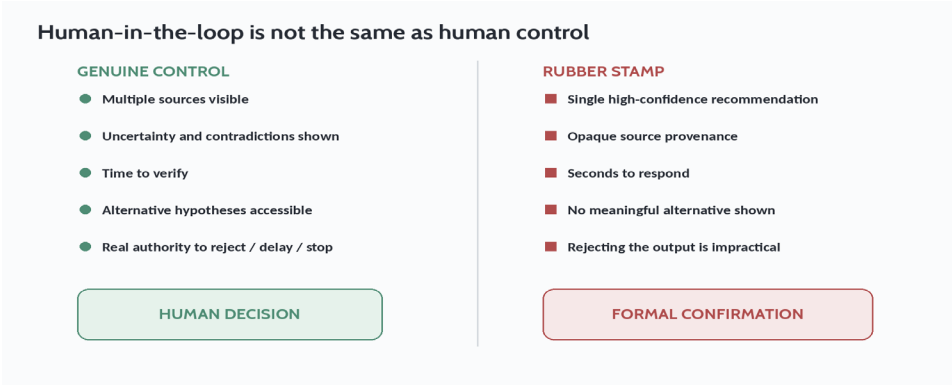


Figure 2. A human can remain physically “in the loop” while the system has already framed the decision so tightly that independent judgement becomes largely formal.

02 DIGITAL TARGETING INFRASTRUCTURE: PALANTIR AS AN ILLUSTRATIVE CASE

Publicly documented capabilities are used to show workflow structure, not to allege autonomous targeting

Evidence discipline: what public sources show - and what they do not

DOCUMENTED IN PUBLIC MATERIALS	NOT ESTABLISHED BY THOSE MATERIALS
Target Boards and High Priority Target Lists exist in the documented Target Workbench data model. [24,25]	That a Palantir product independently authorised lethal force in a specific operation.
Target intelligence can include SIGINT, OSINT, IMINT, ELINT, HUMINT, GEOINT and confidence/source fields. [23]	The degree of automation or model autonomy in a specific classified deployment.
Palantir and Anthropic announced Claude availability for U.S. defence and intelligence organisations through AIP on AWS. [27]	That Claude, Palantir or another AI system caused the 2026 Iran incidents discussed by the UN fact-finding mission. [29]

In its 2025 annual report Palantir describes Gotham as a platform that fuses data from multiple domains and sensors almost in real time - from operations centres down to the tactical level. According to the company, the system improves situational awareness, accelerates operational decision-making and enables the transfer of analytical work to operational users involved in planning and executing missions. AIP, in turn, is described as an integrated environment for using AI models and large language models, including decision interfaces, feedback loops, task hand-offs between AI agents and humans, audit tools, access controls and human-review checkpoints. [\[22,26\]](#)

Public documentation of Target Workbench makes this architecture concrete. It provides for Target, Target Board and High Priority Target List objects; intelligence from the domains SIGINT, OSINT, IMINT, ELINT, HUMINT, GEOINT and others can be attached to a target. The data structure also includes location, information source and a confidence indicator. This confirms the existence of a software model in which an object can be represented as a target, linked to intelligence and placed in a prioritised workflow. [\[23,24,25\]](#)

The link between Palantir's infrastructure and modern language models is not merely theoretical. In November 2024 Palantir, Anthropic and Amazon Web Services announced the provision of Claude models to U.S. defence and intelligence organisations through Palantir AIP. According to The Wall Street Journal of 28 February 2026, U.S. Central Command used Claude for intelligence assessment, target identification and the modelling of combat scenarios. The published information does not, however, disclose the technical configuration of any concrete operation and does not prove that Claude was employed specifically through the Palantir platform. [\[27,28\]](#)

The 17 September 2026 report of the Independent International Fact-Finding Mission on Iran found reasonable grounds to believe that two U.S. airstrikes on clearly identifiable civilian objects were indiscriminate and amounted to war crimes. The Mission did not attribute those strikes to Palantir, Claude or any other specific AI system. The relevance for this report is narrower: acceleration cannot repair stale, incomplete or inadequately verified source data; a digital decision chain can instead propagate such weaknesses more quickly toward the use of force. [\[29\]](#)

Such integration does not necessarily transfer to a machine the formal right to decide on the use of force. But it can structure the conditions under which a human takes that decision: determining which objects enter the field of attention, which pieces of information are linked together, what is labelled a priority and how much time remains for verification. Algorithmic influence can therefore become substantial long before autonomous weapons appear and even while a human formally remains in the decision loop.

It is noteworthy that Palantir itself acknowledges in its annual report that there are decisions whose consequences are too complex or too significant to automate or to accelerate excessively, even with human participation. The company also states that decisions affecting fundamental rights should not be left entirely to computers. These statements are not an independent guarantee of the platform's safety, but they confirm the very framing of the problem: the presence of a human does not eliminate risk if the architecture of the system has already constrained the space of judgement.

The question of this report is therefore not "Is Palantir dangerous?" but rather: what technical, legal and organisational constraints must apply to any platform that combines intelligence data, models, target objects, priorities and operational workflows. Discussion of those constraints should begin before such infrastructure becomes an indispensable part of the command loop.

A dangerous form of delegation can begin long before an automatic shot is fired - at the moment when a software system acquires the practical ability to determine what a human will see as a target, which data he or she will trust and which alternatives there will still be time to consider.

03 SPEED, UNCERTAINTY AND THE BOUNDARY OF HUMAN DECISION

Return for a moment to the hypothetical example in section 2. Not every recognition system is probabilistic. Yet many machine-learning systems produce confidence, risk or similarity scores. A value such as 94 or 96 percent does not by itself mean that the classification is correct with a probability of 94 or 96 percent. For such an interpretation one would need a validated calibration of the model, known false-positive and false-negative rates, account of base-rate frequencies, a description of the test set and proof that the real conditions of use match the conditions of testing. Even a well-calibrated model can lose reliability when the environment, adversary behaviour or the quality of input data changes. ^[30]

Uncertainty does not arise only inside the model. It can result from a sensor error, unclear data provenance, an outdated record, the incorrect merging of information about different objects, missing context, deliberate camouflage or the substitution of input data. A technically sound model can therefore produce an operationally wrong recommendation on the basis of unreliable or outdated information. A reliable system must display separately the uncertainty of the data, the uncertainty of the model and the uncertainty of the legal or operational assessment - rather than conceal them behind a single numerical indicator.

International humanitarian law likewise cannot be reduced to a binary classification “military object – civilian object”. Determining a military objective, assessing the expected military advantage, possible collateral damage and feasible precautions all require contextual judgement. A software system can systematise facts, flag contradictions and signal the absence of required data, but it must not replace the responsible human determination of the lawfulness of a concrete attack.

An increase in potential harm therefore requires more than a raised confidence threshold of the model. What is needed is multi-level escalation logic: confirmation from an independent source or sensor, verification of the currency of the data, analysis of alternative identifications of the object, assessment of possible civilian presence, elevation of the authorisation level and, where practicable, the granting of additional time for verification. If critical uncertainty remains unresolved, the system must refuse to recommend, switch to purely analytical support or block the transmission of a decision into the effector loop. A higher numerical threshold without these measures can create not safety but a technically packaged illusion of certainty.

A genuine human decision exists only when the authorised person has sufficient time, the necessary information, professional competence and the real power to reject, postpone or terminate the system’s action. The operator must see not only the final recommendation but also the provenance of the data, material contradictions, the limitations of the model and alternative assessments. An interface that displays a single priority answer while simultaneously creating time pressure can reinforce the human tendency to accept the automated recommendation. In that case the human in the loop becomes a formal approver - a presence that creates the appearance of control without delivering its substance.

Speed can be a legitimate operational advantage. But it must not be purchased by eliminating precisely those conditions that make a decision human: time for doubt, the possibility of objection and the right to stop the application of force.

04 ENGINEERING RESTRAINT UNDER UNCERTAINTY

The report uses “technical empathy” only as a normative metaphor for verifiable system restraint

In the literal sense the answer is no. A machine does not experience compassion, does not recognise human dignity and does not bear moral responsibility. The notion of “technical empathy” used in this report is therefore not an established scientific or legal term. It is a deliberately chosen normative designation for an architecture that converts doubt, the protection of human life and the limits of permissible action into verifiable system behaviour.

The first element of such an architecture is the capacity to refuse an output. If the status of a person or an object has not been confirmed, the absence of information must not be interpreted as proof that an attack is permissible. If sources contradict one another, data are outdated, their provenance cannot be verified or the model is operating outside its validated conditions of use, the system must flag the uncertainty, request additional verification or decline to generate a recommendation that could lead to the application of force.

This rule can be called a presumption in favour of preserving life under critical uncertainty. It is offered here as a normative design principle, not as a free-standing norm of existing international humanitarian law. Its technical meaning is that the state “I do not know” is recognised as a permissible and necessary system outcome. In a high-risk loop the ability to refuse an answer on justified grounds may be a more important quality indicator than the ability to produce an answer in every case.

The second element is the technical representation of protected statuses and constraints. Geographic zones, lists of medical and cultural objects, temporal restrictions, indicators of civilian presence, prohibited target classes and requirements for additional authorisation can form an external protective layer. The model must not have the authority to cancel or rewrite such constraints on its own.

These mechanisms cannot, however, fully determine whether an object is a military objective, whether the principle of proportionality is observed or whether feasible precautions have been taken. They enforce pre-formalised minimum boundaries; the final contextual legal assessment remains the duty of an authorised human.

“Technical empathy” has practical meaning only when it is expressed in measurable and verifiable requirements. The conditions under which the system must refuse an answer, the grounds for escalating a decision to a higher level, the requirements for independent confirmation, the maximum permissible age of data, the minimum information to be provided to the operator, the permitted modes of safe degradation and the authorities for terminating action must all be defined in advance. These mechanisms must be tested, logged and independently audited. Without that, the word “empathy” remains a rhetorical metaphor that creates no real constraint.

Uncertainty does not always require the complete shutdown of every function. A safe reaction depends on the task, the nature of the threat and the possible consequences of inaction. A system may continue observation, warning, navigation or protective functions while simultaneously blocking the transition to lethal action. But if the factual or legal foundations required for an attack are absent, speed must not turn a lack of evidence into permission to apply force.

Such an approach does not make the machine a moral subject. It makes human constraints part of the technical architecture. “Technical empathy” begins not when a machine claims to understand human suffering, but when the system cannot conceal critical doubt and cannot push an insufficiently grounded decision through to the application of force. A military system that can accelerate the engagement of a target but cannot stop before an unproven target demonstrates not a higher form of intelligence but a deficit of engineering and political control.

DESIGN TRANSLATION

The engineering meaning of “I do not know” is not passivity. It is a defined system state that can trigger additional verification, higher authorisation, advisory-only mode or an effector block.

05 A THEORETICAL ARCHITECTURE FOR SAFE GOVERNANCE OF MILITARY AI

An external safety envelope instead of trust in the model itself A system that generates a high-risk recommendation must not be the sole controller of its own correctness. A model can err, operate outside its validated conditions of use, receive compromised data or be altered after deployment. Between the analytical AI component and the effector loop there must therefore stand an external safety envelope - a set of technical and organisational mechanisms that verify the fulfilment of pre-defined conditions of admission.

That envelope does not determine whether a decision is morally or legally correct. It checks only formalisable preconditions: the permissible version of the model and policy, user authorities, the integrity of software components, the provenance and currency of required data, the presence of mandatory independent confirmations, geographic and temporal restrictions, special categories of objects and the established order of authorisation. If a mandatory condition is not met, the recommendation must not be passed into the force channel.

The analytical model must not possess direct authorities, credentials or an interface with which to bypass this envelope. Every critical action must pass through a single controlled checkpoint.

An architecture in which the model is technically able to circumvent constraints does not become safe merely because an operator is present at the end of the process.

A separate software component is not yet independent. The main model and the protective envelope may share the same sensors, libraries, cloud infrastructure, time services, keys, communication channels or supply chains. Independence must therefore be demonstrated by an analysis of trust boundaries, separation of authorities, least privilege, protected channels, preferably diverse implementations and testing for common-mode failures. Independence is a verifiable property of the architecture, not the name of a component.

Policy-as-code as the technical execution of a limited set of rules Part of the constraints can be expressed as formalised rules executed by a separate control mechanism (policy as code). In this way one can lock in geographic prohibitions, temporal windows, permitted modes, a minimum set of sources, the maximum age of data, restrictions by object category, role constraints and the order of authorisation.

Every rule must have an established source of authority, a defined scope of application, an owner, a version number, an entry-into-force date, a set of tests and a procedure for resolving conflicts. Changes must undergo separate authorisation, be signed, logged and verified before deployment. The main model must not be able to alter the rules, the exceptions or the conditions of its own activation.

Formalised rules in code can cover only a limited and pre-defined portion of requirements. They cannot fully automate the assessment of military advantage, proportionality, doubts about the status of an object or feasible precautions. Those questions depend on concrete context and require human legal and command judgement.

The system must therefore make explicit not only the rules that have been applied but also the domain they do not cover. Technical passage of all machine checks does not turn an attack into a lawful one. Otherwise automation would create a dangerous form of normative laundering: a decision would appear permissible merely because the software had found no prohibition.

Human decision as authority, not as an interface action Real human control requires that the authorised person possess sufficient time, competence, information and the real power to reject a recommendation. The interface must display the provenance of key data, their currency, material contradictions, the limitations of the model, alternative assessments, the grounds for the recommendation and the possible consequences of error. A final confidence score without this context is insufficient.

For certain high-risk actions a dual-authorisation principle may be applied. Multi-factor authentication and dual authorisation perform different functions. Multi-factor authentication confirms the identity of one user. Dual authorisation requires the coordinated decision of two authorised and organisationally separated roles. The two mechanisms may be used together; they are not interchangeable.

Two confirmations do not amount to two independent decisions if both persons receive the same opaque recommendation, operate under the same time pressure and lack the possibility of checking the original data. The distribution of authorisation can reduce the risk of a single error and document the participation of specific individuals, but it does not by itself determine legal responsibility. Responsibility depends on authority, actual knowledge, command relationships and applicable law.

The expiry of a time limit must not automatically constitute consent to the application of force. For an offensive lethal action the absence of the required confirmation must mean the absence of permission. The concrete reaction of the system to a timeout or loss of communication must, however, be determined by a prior hazard analysis, because for platform protection, collision avoidance or the defeat of an immediately approaching threat a complete stop can itself create risk.

Runtime safety monitoring and the limits of automatic observation Verification before deployment does not guarantee unchanged behaviour during operation. Updates, new data, changes in the environment, adversary action, sensor failure or a cyber attack can all affect the system. Runtime monitoring is therefore required that observes the state of the model, the data, the rules and the interfaces.

Monitoring can detect an unexpected change of model version, a breach of integrity, sensor disagreement, anomalous request frequency, a shift in the distribution of results, an unusual rise in positive classifications, the loss of a mandatory source or an attempt to access a prohibited interface. These events must be compared with established operational boundaries.

A monitoring signal does not prove error. A change in data distribution may indicate model failure, adversary manipulation or a real change in the military situation. In conflict, reliable ground-truth data are often absent. Moreover an adversary can deliberately create anomalies in order to trigger automatic blocking and turn the protective mechanism into a denial-of-service tool.

Reactions must therefore be graduated: warning to the operator, request for an additional source, restriction of individual functions, elevation of the authorisation level, transition to advisory mode, transfer of control to a human or blocking of a specific effector channel. A single unexplained signal must not automatically shut down the entire system. Monitoring is a source of evidence for a decision, not an independent instance of truth.

The right of technical intervention must be embedded in a pre-approved governance structure. Authorised roles, stop criteria, the required evidentiary threshold, the escalation order, protected authentication of the command, conditions for emergency override of a block and mandatory subsequent review must all be defined. Otherwise the safety mechanism itself can become a means of abuse, sabotage or unauthorised interference.

Safe degradation instead of universal shut-down There is no universal safe state for every military AI system. A complete stop may be correct for an offensive force channel but dangerous for navigation, collision avoidance, attack warning or protection against an immediately approaching munition. A safe reaction must therefore be defined for each function separately on the basis of an analysis of the consequences of action and of inaction.

The architecture must provide for pre-validated degradation modes: continued observation without the transmission of commands, transition to purely advisory mode, blocking of weapons employment, manual control, a limited protective function, termination of the task, return of the platform or transition to another validated mode with a minimal risk level. The choice must not be improvised by the model itself after a failure has occurred.

Return from a restricted mode requires renewed integrity verification, confirmation of permissible versions, restoration of trusted channels and fresh authorisation. The reason for degradation, the system's actions, human interventions and the grounds for restoration must be preserved for subsequent investigation.

The goal of such an architecture is not to create an error-free machine - that promise would be technically untenable. The goal is to ensure that an error, an attack or uncertainty cannot pass unnoticed through the entire chain and turn into irreversible application of force. Safe governance begins where the system not only knows how to act but is also technically deprived of the ability to act beyond the authorities granted to it.

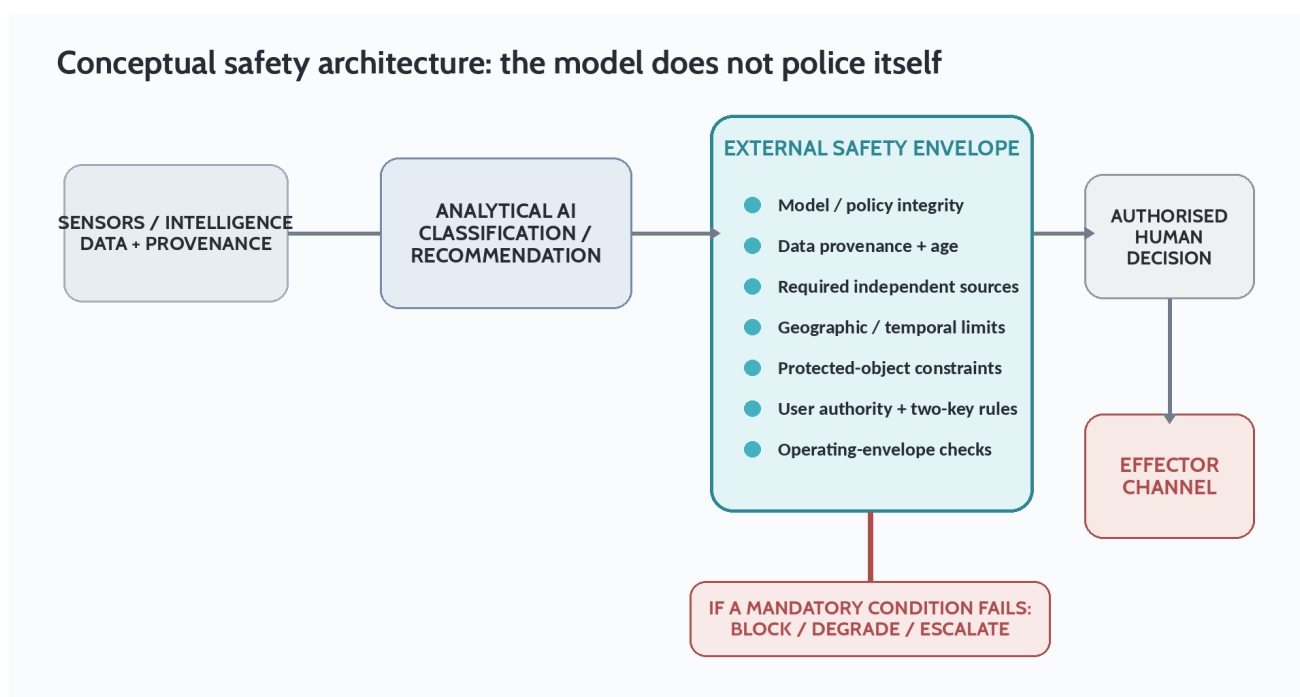


Figure 3. Conceptual external safety envelope. The control layer verifies formalised preconditions but does not replace contextual human legal judgement.

06 CYBERSECURITY, OBSERVABILITY AND INDEPENDENT OVERSIGHT

The expanded attack surface of military AI Traditional threats include account compromise, malicious updates, substitution of software components, breaches of network segmentation, attacks on communication channels and dependence on external suppliers. AI-specific risks include poisoning of training and operational data, compromise of model parameters, adversarial inputs, sensor spoofing, manipulation of external sources, extraction of confidential information and attempts to drive the system outside its validated operating envelope. In systems that use large language models there are additional risks of prompt injection, unsafe use of connected tools and the transmission of malicious instructions through external documents or information streams.

Particular danger attaches to the compromise of trust rather than mere availability. A system may continue to function technically while using a substituted model, an outdated policy or a manipulated source. Such an incident is harder to notice than a complete failure: outwardly the system remains operational, yet its recommendations are already formed on an unreliable basis.

It is therefore necessary to verify the provenance of models and data, digital signatures of artefacts, version hashes, the composition of software dependencies, authorities for updates, the integrity of configurations and the permissibility of connecting external sources. NIST SP 800-218A can serve here as a technical orientation for the secure development and integration of generative AI and dual-use foundation models. It is not, however, a specialised standard for the operation of military systems and does not by itself establish mandatory requirements for national-security systems.

Observability and incident response Protection requires observability at four levels simultaneously: infrastructure, data, model and decision process. Changes to software components, network events and user actions must be recorded, as must the provenance of input data, versions of models and rules, shifts in result statistics, cases of model refusal, divergence between sources and attempts to circumvent established constraints.

It is not necessary to create a separate organisation called an “AI-SOC”. That term does not yet denote an accepted institutional standard. A more realistic approach is to combine specialised AI competences with existing SOC, SIEM, MLOps, system-security, functional-safety and incident-response functions. In this report AI-security monitoring should be understood as a functional model, not as a mandatory name of a unit.

A detected anomaly must not automatically be treated as proof of attack or failure. A change in system behaviour may be caused by compromise, sensor error, weather conditions, new adversary tactics or an objective change in the operational environment. Moreover an adversary can deliberately create anomalous signals in order to trigger protective blocking and turn monitoring into a means of denial of service.

Response must therefore be a sequence: detection of the signal, verification of its quality, comparison with independent sources, assessment of possible consequences and selection of a proportionate restriction mode. Depending on the risk this may be a warning, additional verification, restriction of individual functions, transition to advisory mode, human control or blocking of a specific effector channel.

Forensic traceability and protection of logs For a high-risk decision a sufficient volume of evidence must be retained to establish which data, models, rules and human actions influenced the result. This includes identifiers and versions of the model, configurations and policies, provenance of key data, temporal parameters, material input signals, the system's result, triggered constraints, operator actions, authorisations, mode changes and subsequent updates.

The goal is not mandatory bit-level reproducibility of every event, which may be impossible because of streaming data, external sources, non-deterministic computation and changing infrastructure. What is required is risk-oriented forensic reconstructability: the retained data must be sufficient to restore the essential course of the decision, to verify compliance with established procedures and to determine the contribution of the system and of the persons involved.

An append-only log by itself does not guarantee protection against deletion, substitution or deliberate non-recording of events. An administrator or an attacker can delete the entire log, replace the history, alter timestamps or prevent an event from being recorded before it is written. Cryptographic chaining of successive records, digital signatures, protected key management, trusted time synchronisation, separation of access rights, external or backup storage and regular completeness checks are therefore required.

At the same time such logs may contain information about intelligence sources, targets, sensor capabilities, rules of engagement and personal data. Completeness of logging must therefore be combined with classification of information, minimisation of access, segmentation of storage, established retention periods and a protected procedure for making materials available to authorised reviewers. Audit must not create a new intelligence vulnerability.

Independent oversight and distribution of authority Independent oversight cannot be secured merely by creating a separate software component or unit. What is required is organisational separation of interests, direct access to technical evidence and the authority to report a critical risk outside the operational chain that has an interest in continued use of the system. The oversight function must not, however, independently replace the commander or acquire unlimited power to manage the application of force.

Authorities to restrict the system must be distributed in advance among operational command, cybersecurity and system-safety functions, technical specialists and legal advisers. For each level of intervention the grounds, evidentiary threshold, authorised roles, dual-authorisation order, protected command channel, conditions for emergency override and mandatory subsequent review must be defined.

A stop command is itself a critically important object of protection. It must undergo strict authentication, be protected against replay and forgery, be logged independently of the main system and not depend on a single administrator, device or communication channel. Otherwise a mechanism intended to prevent uncontrolled action can become a tool of sabotage or unauthorised interference.

Independence of oversight does not mean absence of connection with military command; it means the impossibility of concealing a critical risk, altering evidence or continuing to employ the system without the established procedure. A State cannot invoke the complexity of an algorithm as justification for the absence of evidence. A system capable of participating in the application of force yet unable to establish how a decision was formed may remain operationally effective - but it ceases to be a responsible instrument of State power.

SECURITY OBJECTIVE

A high-risk system must be able to prove which model, policy, data and human authorities shaped a critical decision. Availability alone is not enough; integrity and forensic reconstructability are central.

07 ASSURED INTERRUPTION OF THE FORCE-APPLICATION LOOP

“Right to disconnect” is used here as shorthand for a pre-defined technical and organisational capacity to withdraw effector authority

The right to disconnect as a condition of retaining control In this report the expression “right to disconnect” is used not as the designation of an existing universal subjective right codified in an international treaty, but as a normative and technical principle. It means a pre-defined competence and a practically realisable technical possibility to deprive an AI system, in a timely, selective and verifiable manner, of the authority to initiate, recommend or enable the application of force.

This does not necessarily mean shutting down the entire platform. The key task is to sever the specific functional link between a result produced by AI and an effector loop capable of causing physical harm. Navigation, communications, observation, crew protection and data collection may continue if the continuation of those functions does not create additional risk.

Such a right must exist not only in regulations. It must be implemented in system architecture, command procedures, the distribution of authorities and the training of personnel. If a State is unable to withdraw the system’s force-application authority in a timely and verifiable way, it is transferring not merely a technical function but a portion of power that it can no longer reliably control.

Multi-level interruption of the effector loop The idea of a single “red button” creates a false sense of security. A centralised disconnect command itself becomes a critical single point of failure and an attractive target for cyber attack. An adversary may attempt to forge such a command, replay a previously intercepted signal, block the control channel or, conversely, shut the system down at the moment when its protective function is needed.

The mechanism for terminating the application of force must therefore be multi-level and admit at least the following forms of intervention:

1. Transition of the AI to an information-and-analysis support mode with no ability to transmit commands to effector systems;
2. Blocking of the interface between the AI component and the force-application loop;
3. Revocation of the digital authorities, keys or tokens required to form or execute a command;
4. Isolation of a defined network segment or trusted control channel;
5. Deactivation of a specific weapons function while preserving the remaining safe functions of the platform;
6. Local emergency intervention on the platform itself, with priority over remote control.

The stop mechanism should, as far as possible, be independent of the model it controls. The AI must not be able to alter the criteria of its own deactivation, disable the control component or restore a blocked effector channel.

At the same time a complete stop is not in every case the safest solution. For missile-defence systems, active platform protection, collision avoidance or the maintenance of life-critical functions an instantaneous cessation of operation can itself increase danger.

A safe state must therefore be defined for each function on the basis of a prior analysis of threats and consequences. Depending on the situation it may mean a stop, limited continuation, transition to direct human control, return of the platform or retention of protective functions only.

Grounds and authorities for a stop Criteria for intervention must be established before the system is deployed, not improvised during an operation. Grounds for interrupting the effector loop may include, inter alia:

- breach of the integrity of software or of a signed employment policy;
- detection of an unknown or unauthorised model version;
- loss of confirmed provenance of critical data;
- behaviour of the system outside its validated operational envelope;
- attempt to circumvent established constraints;
- loss of required sensor data or material contradictions among them;
- impossibility of ensuring the required level of human control;
- indications of compromise of the model, the control infrastructure or the effector interface.

A technical anomaly by itself does not always prove that the system has been compromised. An adversary can deliberately create anomalies in order to provoke repeated stops and thereby turn the safety mechanism into a means of denial of service. The reaction must therefore be graduated: from warning and restriction of individual functions to immediate blocking of a specific force-application channel.

Automatic stop is permissible for a narrow circle of pre-defined and unambiguously verifiable events - for example, a breach of the cryptographic integrity of a mandatory policy or the loss of a trusted authorisation component. In ambiguous cases the decision must be accompanied by human verification and an established escalation procedure.

Authorities must be distributed among the operator, command, security specialists and other responsible functions. A cybersecurity unit may be given the right to initiate an immediate technical block upon detection of a critical threat. That must not, however, turn it into an independent body that decides on the application of force or on the resumption of combat mode.

For especially critical actions a two-key authorisation is advisable: the decision is confirmed by two independent authorised persons or functions. The two-key principle must not be confused with multi-factor authentication. Multi-factor authentication confirms the identity of one subject; a two-key procedure divides the authority among several subjects.

The stop channel itself must be protected against forgery, replay of commands, blocking of communications and unauthorised restoration of the system. Independent control channels, cryptographic verification of commands, control of their currency, logging and regular testing of procedures under realistic conditions are required.

Resumption of operation after a stop A stop of the system must not end with an automatic return to the previous mode. Restoration of effector authorities is a separate decision that requires stricter verification than an ordinary technical restart.

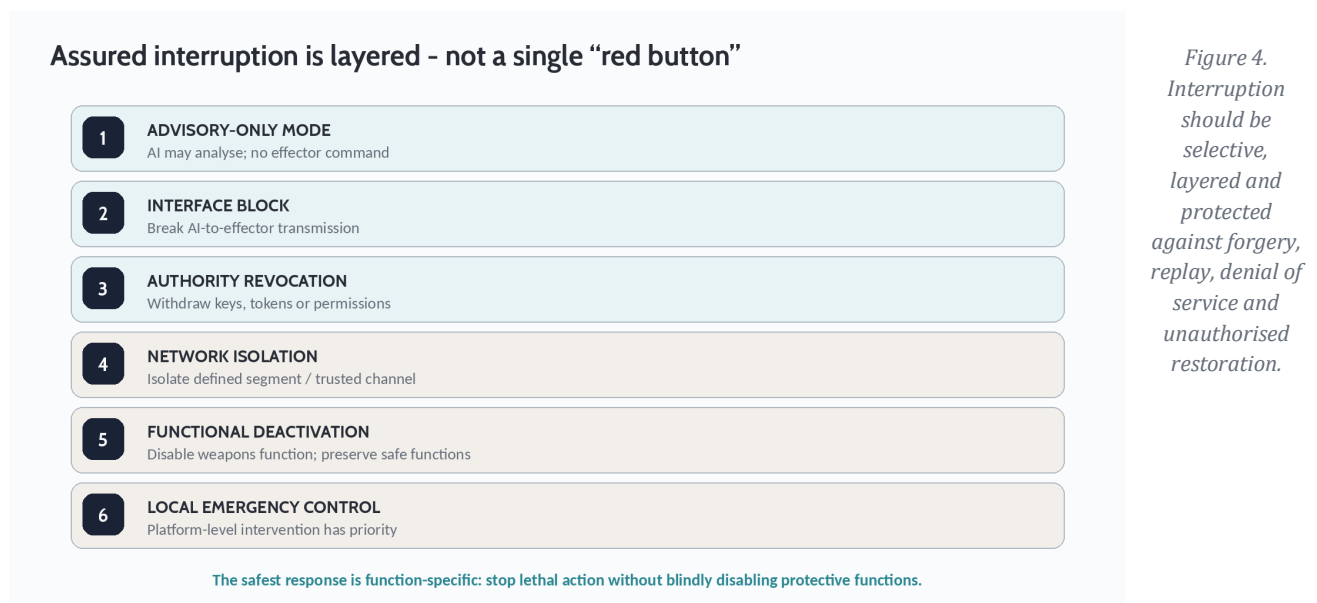
Before re-enablement the reasons for the protective mechanism's activation must be established, digital evidence preserved, versions of the model and software components verified, the integrity of policies and data confirmed, and the possibility of a common-mode failure in related systems excluded. After a serious incident an independent technical assessment and documented confirmation of the system's readiness for further operation must be conducted.

Resumption of a function capable of leading to the application of force must not be carried out by the same subject that unilaterally initiated its stop or performed the remedial work. Separation of roles reduces the risk of an erroneous, hasty or interested restoration.

U.S. Department of Defense Directive 3000.09 embeds the principle of governability: AI-enabled capabilities must be designed and developed so that they perform their intended functions, detect and prevent unintended consequences, and allow deployed systems exhibiting unintended behaviour to be stopped or deactivated. This does not establish a universal architecture and does not itself create an international legal norm, but it confirms the fundamental importance of governability and of the possibility of intervention. [6]

ICRC positions likewise emphasise the need for effective human control and the ability to intervene in a timely manner in the operation of autonomous weapon systems. These positions should be correctly presented as normative recommendations of an authoritative humanitarian organisation, not as an undisputed statement of already binding universal law. [3]

The right to disconnect should therefore be regarded not as an additional safety feature but as a technical condition for the preservation of human and State authority over the application of force. A system whose effector authorities cannot be withdrawn in a timely, selective and verifiable manner must not be admitted to an autonomous force-application loop.



08 ACCOUNTABILITY, INTERNATIONAL HUMANITARIAN LAW AND THE OSCE

Accountability does not disappear, but establishing it becomes more difficult. The principal risk of military AI is not the literal disappearance of a subject of responsibility. An algorithm does not become an independent legal subject, and State and human responsibility do not cease merely because an automated system is used. The distribution of functions among developers, suppliers, military command, operators, data specialists and cybersecurity units can, however,

substantially complicate the determination of who possessed the necessary information, authorities and practical possibility of preventing an unlawful consequence.

Several levels of responsibility must be distinguished:

- the international legal responsibility of the State;
- the individual criminal responsibility of specific persons;
- command responsibility, taking account of actual control, knowledge and the applicable legal standard;
- disciplinary and administrative responsibility;
- institutional accountability for the procurement, testing, introduction, updating and operation of the system.

The existence of responsibility at each of these levels is established by different legal criteria. It depends on the role of the person, his or her authorities, the information available, the nature of the violation, the degree of control and the concrete contribution to the decision. The participation of two persons in an authorisation procedure therefore does not automatically mean a division of legal responsibility between them.

Two-key authorisation, logging and the recording of human interventions have primarily organisational and evidentiary significance. They make it possible to establish who participated in the decision, what data they possessed and what actions they took. The final legal qualification cannot, however, be determined by a single technical record.

An algorithm cannot serve as a legal shield between the application of force and the responsible human. The answer “the system decided” must be treated not as an exoneration but as evidence that the State permitted the application of force through a mechanism whose decisions it could not adequately explain, verify or stop.

International humanitarian law applies to people and States The use of AI does not alter the obligations of parties to an armed conflict under international humanitarian law. In the selection of targets and the conduct of attacks the principles of distinction, proportionality and the taking of feasible precautions must, inter alia, be observed.

AI can process intelligence data, detect objects, forecast behaviour or propose courses of action. Yet a classification generated by a system is not itself a legal determination of a military objective. Likewise a probabilistic estimate of expected harm does not replace the legal assessment of proportionality.

Such an assessment depends on the concrete operational context: the nature of the object, its actual use, the expected military advantage, possible harm to civilians, the reliability of the data and the availability of alternative means of engagement. A substantial part of these circumstances cannot be reduced to a single numerical indicator or a pre-coded rule.

A technical system can therefore support compliance with individual requirements of international humanitarian law, but it cannot independently confirm the lawfulness of an attack as a whole. The legally significant decision must remain with authorised persons capable of taking account of context, uncertainty and the consequences of erroneous classification.

If the limitations of the system, the quality of the data or the operational situation do not permit the required legal assessment to be carried out, the use of such a system must be restricted or terminated. Commercial secrecy of a supplier, military classification or the technical complexity of a model do not cancel the State’s duty to ensure the lawfulness of the application of force.

It is important to distinguish existing international humanitarian law from proposals for its further development. The ICRC clarifies the applicability of existing norms while at the same time advocating the adoption of new legally binding rules with respect to autonomous weapon systems.

These normative proposals carry significant humanitarian authority, but they must not be presented as already binding universal treaty law.

Legal review of new means and methods of warfare For States Parties to Additional Protocol I, Article 36 establishes the obligation to determine whether the employment of a new weapon, means or method of warfare is prohibited by international law in the study, development, acquisition or adoption of that weapon, means or method.

Not every AI component is automatically a weapon. A State must, however, assess whether a concrete system, the manner of its integration or the intended mode of use constitutes a new weapon, means or method of warfare requiring legal review.

The subject of review must be not an isolated model but the system in its real configuration, including:

- intended scenarios of use;
- the nature and quality of the data employed;
- the degree of autonomy;
- interaction with the operator and command;
- limitations of the operating environment;
- possible failure modes;
- resilience to cyber attack and data manipulation;
- the possibility of cancellation or termination of action;
- the consequences of erroneous classification;
- the measures of human control that are provided for.

Legal review must be conducted before the system is used. It cannot be replaced by performance testing, cybersecurity certification or observation of the system during an operation.

At the same time an initial review must not turn into a permanent licence. A substantial change in the model, data, sensors, level of autonomy, rules of engagement, effector interfaces or intended environment of use may alter the functions and consequences of the system. In such a case it is necessary to determine whether a renewed legal review is required.

Four distinct processes should be kept separate: prior legal review, technical verification, operational authorisation for a concrete employment, and continuous monitoring during operation. Each solves a separate task; none can fully replace the others.

A realistic role for the OSCE The OSCE Code of Conduct on Politico-Military Aspects of Security provides an important political point of reference.

Its paragraph 24 requires participating States to maintain effective guidance and control of their military and paramilitary forces and to ensure protection against accidental or unauthorised use of military means. Paragraph 39 expressly indicates the politically binding character of the Code.

The Code is therefore not an international treaty and does not itself establish detailed legal or technical requirements for military AI systems. Extending the logic of paragraph 24 to algorithmically unforeseen application of force is a reasoned proposal of this report, not an already existing formulation of the OSCE.

The sense of the proposal is this: a State must protect military means not only against unauthorised human access but also against behaviour of an automated system that exceeds the authorised regime, is based on compromised data or leads to an unforeseen transition from analysis to the application of force.

The OSCE should not replace international negotiations on autonomous weapon systems, UN mechanisms or national legal-review procedures. Nor should it become a technical certification body for concrete models. Its comparative advantage lies in politico-military dialogue, conflict prevention and the development of confidence-building measures.

A practical OSCE agenda could include:

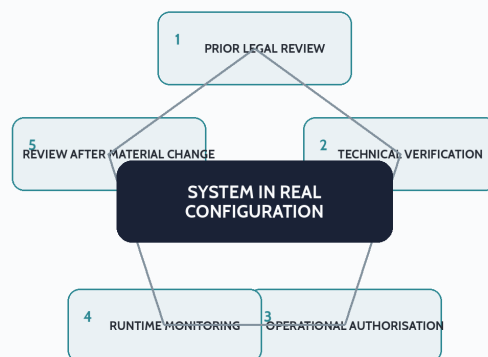
- agreement on shared terminology in the field of military AI;
- voluntary exchange of information on national procedures for legal review and human control;
- establishment of contact points for serious incidents involving military AI systems;
- development of compatible formats for reporting incidents without disclosing operationally sensitive data;
- exchange of experience in testing, safe degradation and the withdrawal of effector authorities;
- expert and command-post exercises on scenarios of erroneous or compromised automation;
- the formation of voluntary transparency and confidence-building measures.

Disclosure of information must be multi-level. Basic principles of control and national procedures may be communicated publicly. More detailed information may be discussed in protected inter-State consultations. Concrete keys, vulnerabilities, network topology and operational stop criteria must remain classified, because their publication can create new opportunities for attack.

OSCE conferences and discussions on artificial intelligence confirm the political relevance of the topic and the existence of space for further dialogue. Without an official decision or an agreed outcome document, however, they cannot be treated as proof that the OSCE already possesses a concrete policy on autonomous weapons or on mechanisms for disconnecting them. [12,13]

The central problem, therefore, is not whether an algorithm can bear responsibility. It cannot. The question is whether States are able to maintain a continuous and demonstrable chain of human authority - from the development and procurement of a system to the concrete decision to apply force. The more complex the technological chain becomes, the higher the duty of the State to document, explain and legally justify the decision taken - not the lower its responsibility.

Continuous legal and technical assurance cycle



A model update, data-source change, new sensor, new autonomy level or changed effector interface can trigger renewed review.

Figure 5. Legal review, technical verification, operational authorisation and runtime monitoring solve different problems and should remain distinct.

OSCE FIT

The OSCE's comparative advantage is not weapons certification. It is politico-military dialogue, confidence-building, shared terminology and structured exchanges that can reduce misperception and unintended escalation. [11,12,13]

09 RECOMMENDATIONS

A multi-level model: law, command, engineering, procurement and confidence-building

#	POLICY / ENGINEERING PRIORITY	EVIDENCE OF IMPLEMENTATION
1	Differentiate by function and risk	System category, autonomy, reversibility and effector authority documented
2	Make human control verifiable	Time, information, competence and real reject/stop authority demonstrated
3	Maintain continuous review	Legal review, testing, operational authorisation, monitoring and change-triggered re-review
4	Demonstrate protective-layer independence	Separate trust domains, authorities, keys and common-mode-failure analysis
5	Predefine safe degradation	Function-specific fallback states and protected withdrawal of effector authority
6	Integrate AI security into existing assurance	Model/data integrity, SIEM/SOC/MLOps/security engineering and incident response
7	Preserve forensic reconstructability	Versioned models/policies, data provenance, operator actions and effector commands
8	Put assurance rights into procurement	Documentation, change control, independent testing, incident access and exit conditions
9	Use OSCE for confidence-building	Terminology, contact points, voluntary exchange, incident formats and exercises
10	Do not admit systems that cannot be controlled	No independent application of force where scope, review, interruption or reconstruction is impossible

First. Distinguish categories of military AI by function and risk States should differentiate systems of intelligence analysis, object recognition, decision support, target selection, platform control and autonomous weapons employment. Identical requirements for all these systems would be technically disproportionate and legally imprecise.

The level of control must be determined not by the name of the technology but by the possible consequences of its use, the degree of autonomy, the reversibility of the decision, the character of effector authorities and the ability of a human to intervene in time.

The closer a system stands to the direct application of force and the less reversible the consequences of its error, the stricter the requirements must be for verification, authorisation, observability and disconnection.

Second. Embed real human control as a verifiable condition of admission OSCE participating States should politically embed the principle of real human control over systems capable of influencing the application of lethal force. This is a proposal of the report, not an assertion that a single universal definition of such control already exists.

Control must include:

- sufficient time for assessing the situation;
- access to material information and information about uncertainty;
- understanding of the functions and limitations of the system;
- training and real authorities of the operator;
- the possibility of rejecting a recommendation;
- the possibility of restricting or terminating the system's effector authorities;
- protection of the human against pressure under which a machine recommendation is treated as obligatory.

The formal presence of an operator is not sufficient. If a person does not understand the grounds of a recommendation, lacks time to verify it or cannot actually change the result, the procedure merely imitates human control.

Third. Create a continuous cycle of legal and technical review For States Parties to Additional Protocol I of 1977 to the Geneva Conventions of 1949, systems that may constitute a new weapon, means or method of warfare must undergo legal review in accordance with Article 36.

It is necessary to distinguish:

1. prior legal review;
2. technical verification and testing;
3. operational authorisation for a concrete employment;
4. monitoring during operation;
5. renewed review after substantial changes.

An update of the model, a change of data, sensors, effector interfaces, level of autonomy or intended environment of use must trigger a formalised assessment of the need for renewed legal and technical review. A system must not retain its previous clearance merely because its commercial name or principal platform has not formally changed.

Fourth. Require demonstrated independence of the protective envelope For high-risk systems an external protective envelope should be provided that controls the permitted actions of the AI and is capable of blocking the transmission of an inadmissible command into the effector system.

Logical separation from the main model is not sufficient. Independence must be demonstrated by architecture and by testing. The protective envelope should, as far as possible, use separate trusted domains, keys, authorities and protected control channels. Common-mode failures must be examined, including dependence on the same sensors, libraries, time sources, suppliers and communication channels.

An external control layer can technically enforce individual constraints, but it must not be presented as a mechanism for the full automation of international humanitarian law. Contextual assessment of distinction, proportionality and precautions remains the task of authorised humans.

Fifth. Design safe degradation and the withdrawal of effector authorities For every critical function the safest state must be defined in advance, taking account of the nature of the mission and the consequences of a stop. A universal requirement of complete shut-down can be dangerous for missile defence, crew protection, collision avoidance and other functions whose cessation can increase harm.

Depending on the situation safe degradation may mean:

- transition to an information-and-analysis mode;
- prohibition of the transmission of commands to effector systems;
- restriction of autonomy;
- transfer of control to a human;
- retention of navigational, communications or protective functions only;
- return of the platform;
- local deactivation of a specific weapons function.

An unknown model version, a breach of the integrity of a mandatory policy, the loss of a trusted channel or compromise of an effector interface must not automatically authorise further autonomous application of force.

A stop command itself requires protection. Its forgery, replay, blocking and unauthorised lifting of a restriction must be prevented. For certain actions strong authentication of each participant should be combined with two-key authorisation by independent roles.

Sixth. Integrate AI monitoring into the existing security system Military organisations must monitor not only networks and end devices but also the integrity of models, the provenance of data, configuration changes, behavioural deviations and the interaction of AI with effector systems.

These functions can be incorporated into existing SOC, SIEM, MLOps, security-assurance and incident-response structures. The term “AI-SOC” is permissible as the author’s suggestion for designating the aggregate of such functions, but it must not be presented as an already established international standard.

A signal of anomaly or drift must be treated as grounds for verification, not as final proof of compromise. A change in behaviour may be caused by an attack, a sensor fault, a model error or a real change in adversary tactics. Moreover an adversary can deliberately provoke the triggering of protective mechanisms with the aim of denial of service. Reactions must therefore be graduated and accompanied by independent verification.

Seventh. Ensure risk-oriented forensic traceability For high-risk decisions information sufficient for subsequent technical and legal reconstruction must be retained:

- version of the model and its configuration;
- version of the rules then in force;
- provenance of material input data;
- temporal basis of the event;
- results generated by the system;

- indicators of uncertainty;
- constraints that were triggered;
- human interventions and authorisations;
- effector commands that were transmitted.

The requirement must not be reduced to bit-level reproducibility of every event: under real-time conditions and with non-deterministic components that may be unattainable. The goal is sufficient reconstructability of the decision chain.

Logs must be protected by cryptographic integrity, separated access, backup or external storage and protected time synchronisation. At the same time secrecy, protection of personal data and operational security must be taken into account. Logging itself must not create a centralised body of intelligence information accessible to an adversary.

Eighth. Transfer security requirements into procurement and contracts with suppliers A State must not procure a system if it does not obtain sufficient rights and technical possibilities for its verification, restriction and safe operation. Contracts for high-risk systems should provide for:

- documentation of functions and known limitations;
- version and change management;
- notification of critical vulnerabilities and incidents;
- supply-chain control;
- access to data necessary for investigation;
- the right to independent testing;
- rules for updates and re-certification;
- conditions for the termination of support and safe withdrawal from service.

The procurement of an opaque model does not transfer to the supplier the State's responsibility for the application of force. A State may entrust a contractor with the development of a system, but it cannot transfer to that contractor its international legal obligations.

Ninth. Focus the work of the OSCE on realistic confidence-building measures The OSCE should begin not with the development of a detailed technical standard but with politically and institutionally attainable measures. Within the relevant politico-military formats participating States could develop:

- shared terminology;
- voluntary national statements of principles of human control;
- exchange of approaches to legal review and change management;
- contact points for serious incidents;
- compatible formats for incident notification;
- exchange of experience in testing and safe degradation;
- expert and scenario-based exercises;
- voluntary confidence-building measures.

Transparency must be differentiated. Principles and structures of responsibility may be disclosed publicly. More detailed evidence may be made available to States and authorised review bodies in a protected mode. Operational parameters, keys, vulnerabilities and the precise conditions of disconnection must not be published, because that can facilitate the circumvention of protection or the conduct of a cyber attack.

The development of concrete technical specifications is better left to specialised national and international organisations. The role of the OSCE consists first of all in forming a common political language, preventing misperception and reducing the risk of unintended escalation.

Tenth. Establish the principle of non-admission where control is impossible Ethical and legal principles must be translated into verifiable requirements as far as that is technically possible: the states “I am not certain”, “additional verification is required”, “independent authorisation is required” and “the effector channel is blocked”.

Human dignity, distinction, proportionality and the protection of civilians cannot be fully reduced to machine-executable code. Technical rules can constrain a system, but they do not replace human legal and moral judgement.

If a State cannot define the permissible domain of application of a system, ensure legal review, withdraw effector authorities and reconstruct a critical decision, such a system must not be admitted to the independent application of force.

Technological complexity is not a ground for lowering requirements. On the contrary: the less explainable and predictable a system is, the narrower the freedom of action that should be granted to it.

Proposed implementation pathway

The recommendations can be introduced in stages. The sequence below is illustrative: it is intended to prevent governance work from waiting for a single comprehensive technical standard.

STAGE	PRIORITY	PRACTICAL OUTPUT
1 · Establish the baseline	Classify systems by function and risk; define validated operating envelopes and review triggers.	System inventory; legal-review path; named authorities; minimum human-control conditions.
2 · Build demonstrable assurance	Implement independent constraints, safe degradation, protected logging and supplier access rights.	Testable stop/restriction mechanisms; reconstructable evidence; change-control and incident procedures.
3 · Build confidence across institutions	Exercise the controls and exchange non-sensitive governance information with partners.	Contact points; scenario exercises; voluntary transparency; protected incident-exchange formats.

10 CONCLUSION

International humanitarian law, the principles of distinction and proportionality, the duty to take feasible precautions and command responsibility did not arise because war became humane. They express an attempt to set a limit where military effectiveness alone cannot serve as sufficient justification. The appearance of AI does not remove that limit and does not lower the requirements placed on the State. On the contrary: the faster and more complex the decision system becomes, the more convincingly the State must demonstrate that it retains control over it.

The principal danger is not that a machine will one day acquire an evil will. Far more realistic is a situation in which it receives effector authorities before the State has created the legal, organisational and technical mechanisms for constraining them. Then a decision formally taken by a human may in

fact become the result of an opaque recommendation, time pressure, automated classification and institutional dependence on the system.

Morality and international humanitarian law cannot be fully converted into software code. Technical rules can block individual categories of targets, restrict areas and times of employment, verify authorities, demand additional authorisation and terminate the transmission of commands. But they cannot exhaustively determine military advantage, assess all consequences for civilians or replace contextual human judgement.

A safe architecture for military AI must therefore not automate the law but create the conditions for its application by responsible people. It must preserve uncertainty rather than conceal it; give a human time and information rather than only a ready conclusion; constrain the authorities of the model; ensure legal review, observability and forensic traceability; and allow the system's ability to act on the physical world to be withdrawn in time.

Technical control is not an alternative to responsibility. Logging does not make a decision lawful. Two-key authorisation does not automatically distribute legal guilt. An external protective envelope cannot by itself conduct a full assessment under international humanitarian law. An emergency disconnect does not guarantee safety if it can be forged, blocked or used as a means of attack. Each of these mechanisms is necessary in a certain context, but only as part of an overall system of legal, command and engineering accountability.

The central conclusion of this report is that the suitability of military AI must be assessed not only by its capacity to detect a target, shorten decision time or raise operational effectiveness. An equally important criterion is the capacity of the State to constrain the system, to verify the grounds of its output, to prevent an inadmissible action and to establish responsibility after an incident.

If a State is unable to define the permissible domain of application of a system, conduct the necessary legal review, ensure real human control, withdraw effector authorities and reconstruct a critical decision, such a system must not be admitted to the independent application of force. Unexplainability is not a permission. Speed is not a legal ground. Technological superiority does not free anyone from responsibility.

The OSCE is not obliged to wait for a universal treaty before beginning work to reduce these risks. Shared terminology, voluntary transparency measures, protected exchange of information on incidents, national contact points and joint principles of human control can reduce the probability of misperception and unintended escalation. At the same time the OSCE must remain a political platform of trust and dialogue, without replacing international humanitarian law, UN negotiation processes and specialised technical standardisation.

The notion of "technical empathy" used in this report does not mean that a machine is capable of experiencing compassion. It is a deliberate normative metaphor designating a technically implemented capacity to refrain from action, to acknowledge the insufficiency of data, to request additional verification or to block an effector loop. Such a system does not become humane. It merely remains within the limits set by humans.

Before granting artificial intelligence the possibility of bringing closer the moment of a shot, a State must prove that it is capable of preventing that shot. The main question is not whether we shall learn to create machines that can kill faster and more accurately. The main question is whether we shall retain the people and the institutions capable of saying to them: "No."

And perhaps the first technical expression of empathy in a military machine should be not the ability to utter "I understand", but the ability, at the decisive moment, to stop:

"I am not certain. Therefore I do not fire."

ANNEX A · MINIMUM ASSURANCE CHECKLIST

This checklist is not a legal test or certification standard. It translates the report's main proposals into questions that can be used during procurement, review, exercises and pre-deployment assurance.

Legal and operational scope

- ☐ Is the intended use and prohibited use explicitly defined?
- ☐ Has the system been assessed for whether an Article 36 review is required, and has that review been completed where applicable?
- ☐ Is the validated operating envelope documented, including environment, data, sensors and autonomy level?
- ☐ Can the system distinguish technical rule compliance from the separate human legal judgement required for a concrete attack?

Data and model assurance

- ☐ Are provenance, age and integrity of material input data visible and recorded?
- ☐ Are mandatory independent confirmations defined for high-risk recommendations?
- ☐ Are confidence or risk scores calibrated and explained rather than displayed as unexplained certainty?
- ☐ Are model, policy and configuration versions cryptographically controlled and auditable?
- ☐ Are drift, anomalous behaviour and out-of-envelope operation monitored?

Human control

- ☐ Does the operator have sufficient time and access to underlying evidence?
- ☐ Can the operator inspect contradictions, alternatives and model limitations?
- ☐ Does the operator have real authority to reject, delay, escalate or stop?
- ☐ Is absence of confirmation treated as absence of permission for offensive lethal action?
- ☐ Where two-key authorisation is used, are the roles genuinely independent rather than merely duplicated?

Cybersecurity and oversight

- ☐ Is there an independent protective layer or equivalent control function outside the model's own authority?
- ☐ Are stop / restriction channels strongly authenticated and protected against replay, forgery and single points of failure?
- ☐ Are logs protected against deletion, substitution and silent non-recording?
- ☐ Can oversight personnel escalate a critical risk outside the operational chain that has an interest in continued use?

Degradation, incident response and restoration

- ☐ Are function-specific safe states defined before deployment?
- ☐ Can lethal or effector authority be withdrawn without unnecessarily disabling protective functions?
- ☐ Does restoration require fresh integrity checks and separate authorisation?
- ☐ Can a critical decision be reconstructed after the event with sufficient technical and human evidence?

ANNEX B · TERMINOLOGY

TERM	WORKING MEANING IN THIS REPORT
AI-DSS	AI-supported decision-support system: a system that analyses data and supports a human decision rather than independently exercising formal authority to use force.
Autonomous weapon system (AWS)	For this report, the ICRC working definition is used: a weapon system that, once activated, can select and engage one or more targets without further human intervention. [3]
Effector channel / loop	The technical and procedural path by which a recommendation or decision can produce physical action, including weapons employment.
External safety envelope	A concept proposed in this report: a technically and organisationally distinct control layer that checks pre-defined admission conditions before a high-risk output can pass to the effector channel.
Technical empathy	An authorial metaphor for engineering restraint under uncertainty. It does not mean that a machine feels compassion; it means that doubt can trigger refusal, escalation, verification or blocking.
Assured interruption / “right to disconnect”	A pre-defined, protected and practically realisable ability to withdraw, restrict or sever an AI system’s authority to initiate, recommend or enable the application of force.
Safe degradation	A transition to a lower-risk, pre-validated operating mode rather than an undifferentiated shutdown.
Forensic reconstructability	Retention of enough evidence to reconstruct the essential decision chain: model and policy versions, data provenance, material inputs, triggered constraints, human interventions and effector commands.
Automation bias	A tendency to over-rely on automated recommendations, including omission and commission errors when decision aids are imperfect. [31,32]
Model calibration	The relationship between a model’s stated confidence and its empirical likelihood of being correct; a raw confidence score is not automatically a calibrated probability. [30]

ANNEX C

PROPOSED OSCE CONFIDENCE-BUILDING MEASURES FOR MILITARY AI

A practical, voluntary framework for transparency, communication and risk reduction

This annex proposes a limited set of voluntary confidence-building measures for AI-enabled military decision-making. The aim is not to create an OSCE technical certification regime, to disclose operational capabilities, or to replace international humanitarian law and national legal-review procedures. The purpose is narrower: to reduce misperception, improve communication and make the structures of human control, incident response and technical restraint more predictable among participating States.

The proposal follows an established OSCE logic. Existing confidence- and security-building measures use transparency, information exchange, contact points, consultation and practical co-operation to reduce the risk of incidents and misunderstanding. The OSCE's cyber/ICT confidence-building measures provide a particularly relevant precedent: they combine voluntary exchanges, national points of contact and communication channels without requiring States to reveal sensitive technical details. The 2026 Swiss Chairpersonship conference on anticipating technologies also confirmed that artificial intelligence is already part of the OSCE security dialogue.

DESIGN PRINCIPLE Confidence-building should focus less on revealing capabilities and more on revealing the existence of control, communication and incident-management structures.

PROPOSED MEASURES

MEASURE		PRACTICAL CONTENT
1	Shared terminology and national statements	Develop a voluntary common vocabulary for AI-DSS, autonomous functions, effector authority, human control, safe degradation and serious AI-related incidents. Participating States could also publish short national statements describing their general approach to human control and governance, without disclosing operational parameters.
2	National points of contact	Establish designated policy and technical points of contact for serious concerns involving military AI. The function could be modelled on the OSCE cyber/ICT point-of-contact practice and used for clarification, de-escalation and rapid communication when a system malfunction, compromise or unexpected behaviour could have cross-border implications.
3	Voluntary exchange on governance procedures	Exchange high-level information on national processes for legal review, testing, change management, operator training, cybersecurity, incident investigation and the withdrawal of effector authorities. The exchange should concern procedures and control structures, not classified targeting criteria, vulnerabilities, keys or system architecture.
4	Protected incident-notification format	Develop a common voluntary format for notifying other participating States, through protected channels, of serious incidents or near misses where AI-enabled military functions create a material risk of misunderstanding or unintended escalation. Notifications should distinguish confirmed facts from preliminary assessments and avoid operationally sensitive detail.
5	Consultation in cases of unusual or ambiguous activity	Create a procedure for requesting clarification when AI-enabled military activity appears unusual, technically ambiguous or potentially destabilising. The purpose would not be verification of a State's weapons programme, but timely clarification of events that might otherwise be interpreted as deliberate hostile action.

PROPOSED MEASURES — CONTINUED

	MEASURE	PRACTICAL CONTENT
6	Scenario-based exercises	Conduct voluntary table-top and command-post exercises around defined risk scenarios: corrupted or outdated data, sensor spoofing, model drift, compromise of a control component, loss of communication, an unauthorised model version, or a breakdown of meaningful human control. Exercises should test communication and governance procedures rather than classified system performance.
7	Exchange of experience on safe degradation	Share non-sensitive lessons on how systems move from normal operation to advisory-only mode, restricted autonomy, human control or blocked effector authority. The objective is to develop a common understanding that a safe state may differ by function and that the ability to interrupt force application must itself be protected against cyber attack or misuse.
8	Voluntary expert demonstrations and technical dialogue	Where States choose to do so, organise expert briefings or demonstrations of generic safety mechanisms such as auditability, two-key authorisation, protected stop channels, independent monitoring and change-control procedures. Participation should not imply endorsement or certification of a specific system.
9	Annual implementation and lessons-learned dialogue	Use an annual expert-level review to discuss implementation, recurring incident categories, lessons from exercises and possible refinements to the voluntary measures. The focus should remain on reducing misunderstanding and unintended escalation rather than creating a permanent technical inspection regime.

SUGGESTED IMPLEMENTATION SEQUENCE

PHASE 1	Agree terminology; nominate national points of contact; exchange short voluntary descriptions of national governance and human-control principles.
PHASE 2	Test communication channels; develop a protected incident-notification template; conduct scenario-based exercises; exchange good practices on safe degradation and post-incident review.
PHASE 3	Hold a recurring implementation dialogue; refine measures on the basis of practical experience; consider optional expert demonstrations or additional voluntary transparency measures.

SAFEGUARDS

- All measures should remain voluntary and politically negotiated; they should not be presented as an existing OSCE rule on military AI.
- No measure should require disclosure of vulnerabilities, cryptographic material, precise targeting criteria, operational stop thresholds, network topology or other information whose release would create a new security risk.
- Participation in an information exchange, exercise or demonstration should not be treated as certification of a system, evidence of legal compliance or acceptance of a particular technical architecture.
- The framework should complement — not replace — applicable international humanitarian law, national Article 36 reviews where required, UN processes on autonomous weapons and specialised technical standardisation.

INSTITUTIONAL BASIS

1. OSCE, Confidence- and security-building measures: transparency, information exchange, consultation, verification and military co-operation as tools for predictability and risk reduction.
2. OSCE Cyber/ICT Security: 16 politically committed confidence-building measures, including voluntary information exchange, points of contact and communication lines.
3. Swiss OSCE Chairpersonship, “Anticipating technologies – for a safe and humane future”, Geneva, 7–8 May 2026.
4. OSCE Code of Conduct on Politico-Military Aspects of Security, especially paragraph 24 on effective guidance and control and protection against accidental or unauthorised use of military means.

This annex is a proposal by MEcoVEA e.V. for expert discussion. It does not represent an adopted position of the OSCE or any participating State.

SOURCES AND NOTES

Selected primary, technical and peer-reviewed references

1. **International Committee of the Red Cross.** Artificial intelligence in the military domain: ICRC submits recommendations to UN Secretary-General. 17 April 2025.
<https://www.icrc.org/en/article/artificial-intelligence-military-domain-icrc-submits-recommendations-un-secretary-general>
2. **International Committee of the Red Cross.** Frequently asked questions: Artificial Intelligence (AI) in the military domain. 11 June 2026.
<https://www.icrc.org/en/article/faq-artificial-intelligence-in-military-domain>
3. **International Committee of the Red Cross.** Autonomous Weapon Systems and International Humanitarian Law: Selected Issues. Position paper, published 3 March 2026.
<https://www.icrc.org/en/article/autonomous-weapon-systems-and-international-humanitarian-law-selected-issues>
4. **ICRC IHL Database.** Protocol Additional I to the Geneva Conventions, Article 36 — New weapons. 8 June 1977.
<https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-36>
5. **International Committee of the Red Cross.** A Guide to the Legal Review of New Weapons, Means and Methods of Warfare. 2006; ICRC online publication record.
<https://www.icrc.org/en/publication/0902-guide-legal-review-new-weapons-means-and-methods-warfare-measures-implement-article>
6. **U.S. Department of Defense.** DoD Directive 3000.09: Autonomy in Weapon Systems. 25 January 2023.
<https://media.defense.gov/2023/Jan/25/2003149928/-1/-1/0/DOD-DIRECTIVE-3000.09-AUTONOMY-IN-WEAPON-SYSTEMS.PDF>
7. **National Institute of Standards and Technology.** Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. January 2023.
<https://www.nist.gov/publications/artificial-intelligence-risk-management-framework-ai-rmf-10>
8. **National Institute of Standards and Technology.** Secure Software Development Practices for Generative AI and Dual-Use Foundation Models, NIST SP 800-218A. July 2024.
<https://csrc.nist.gov/pubs/sp/800/218/a/final>
9. **National Institute of Standards and Technology.** Security and Privacy Controls for Information Systems and Organizations, NIST SP 800-53 Rev. 5, updated version.
<https://csrc.nist.gov/pubs/sp/800/53/r5/upd1/final>
10. **Boulanin, V.; Peldán Carlsson, M.; Goussac, N.; Davison, N.** Limits on Autonomy in Weapon Systems: Identifying Practical Elements of Human Control. SIPRI / ICRC, June 2020.
<https://www.sipri.org/publications/2020/policy-reports/limits-autonomy-weapon-systems-identifying-practical-elements-human-control>
11. **Organization for Security and Co-operation in Europe.** Code of Conduct on Politico-Military Aspects of Security. 3 December 1994.
<https://www.osce.org/fsc/41355>
12. **OSCE Swiss Chairpersonship.** Swiss OSCE Chairpersonship Conference on Anticipating Technologies opened in Geneva. 7 May 2026; conference held 7–8 May 2026.
<https://www.osce.org/chairpersonship/663922>

13. **OSCE Swiss Chairpersonship.** Switzerland takes over OSCE Chair. 1 January 2026; includes the priority “anticipating technologies — for a safe and humane future”.
<https://www.osce.org/chairpersonship/661444>
14. **United Nations Secretary-General.** Lethal autonomous weapons systems, report A/79/88. 1 July 2024.
<https://documents.un.org/doc/undoc/gen/n24/154/32/pdf/n2415432.pdf>
15. **United Nations General Assembly.** Resolution 78/241: Lethal autonomous weapons systems. 22 December 2023.
<https://documents.un.org/doc/undoc/gen/n23/431/11/pdf/n2343111.pdf>
16. **United Nations General Assembly.** Resolution 79/239: Artificial intelligence in the military domain and its implications for international peace and security. Adopted 24 December 2024.
<https://documents.un.org/doc/undoc/gen/n24/426/98/pdf/n2442698.pdf>
17. **CCW Meeting of High Contracting Parties.** Guiding Principles affirmed by the GGE on emerging technologies in the area of LAWS, Annex III to CCW/MSP/2019/9. 2019. Annex extract hosted by the NATO Cooperative Cyber Defence Centre of Excellence.
https://ccdcoe.org/uploads/2020/02/UN-191213_CCW-MSP-Final-report-Annex-III_Guiding-Principles-affirmed-by-GGE.pdf
18. **CCW Group of Governmental Experts on LAWS.** Chair’s summary — First 2026 session of the GGE on LAWS, CCW/GGE.1/2026/WP.2. 1 April 2026.
[https://docs-library.unoda.org/Convention on Certain Conventional Weapons - Group of Governmental Experts on Lethal Autonomous Weapons Systems %282026 %29/CCW-GGE.1-2026-WP.2.pdf](https://docs-library.unoda.org/Convention%20on%20Certain%20Conventional%20Weapons%20-%20Group%20of%20Governmental%20Experts%20on%20Lethal%20Autonomous%20Weapons%20Systems%20-%202026%20-%20CCW-GGE.1-2026-WP.2.pdf)
19. **NATO.** Summary of NATO’s Revised Artificial Intelligence Strategy. 10 July 2024.
<https://www.nato.int/en/about-us/official-texts-and-resources/official-texts/2024/07/10/summary-of-natos-revised-artificial-intelligence-ai-strategy>
20. **U.S. Department of State.** Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy. November 2023. Archived text.
<https://2021-2025.state.gov/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy>
21. **Republic of Korea, Ministry of Foreign Affairs.** Outcome of Responsible AI in Military Domain (REAIM) Summit 2024; “Blueprint for Action”. 10 September 2024.
https://www.mofa.go.kr/eng/brd/m_5676/view.do?seq=322676
22. **Palantir Technologies Inc.** Annual Report on Form 10-K for fiscal year ended 31 December 2025. 2026 filing.
<https://investors.palantir.com/files/2025%20FY%20PLTR%2010-K.pdf>
23. **Palantir Technologies.** Gotham API — Create Intel on a Target / Target Workbench. Accessed September 2026.
<https://www.palantir.com/docs/gotham/api/target-workbench/intel/create-target-intel>
24. **Palantir Technologies.** Gotham API — Create a Target Board. Accessed September 2026.
<https://www.palantir.com/docs/gotham/api/target-workbench/targetBoard/create-target-board>
25. **Palantir Technologies.** Gotham API — Create a High Priority Target List. Accessed September 2026.
<https://www.palantir.com/docs/gotham/api/target-workbench/high-priority-target-list/create-hptl>
26. **Palantir Technologies.** AIP for Defense. Accessed September 2026.
<https://www.palantir.com/platforms/aip/defense/>
27. **Palantir Technologies.** Anthropic and Palantir Partner to Bring Claude AI Models to AWS for U.S. Government Intelligence and Defense Operations. 7 November 2024.

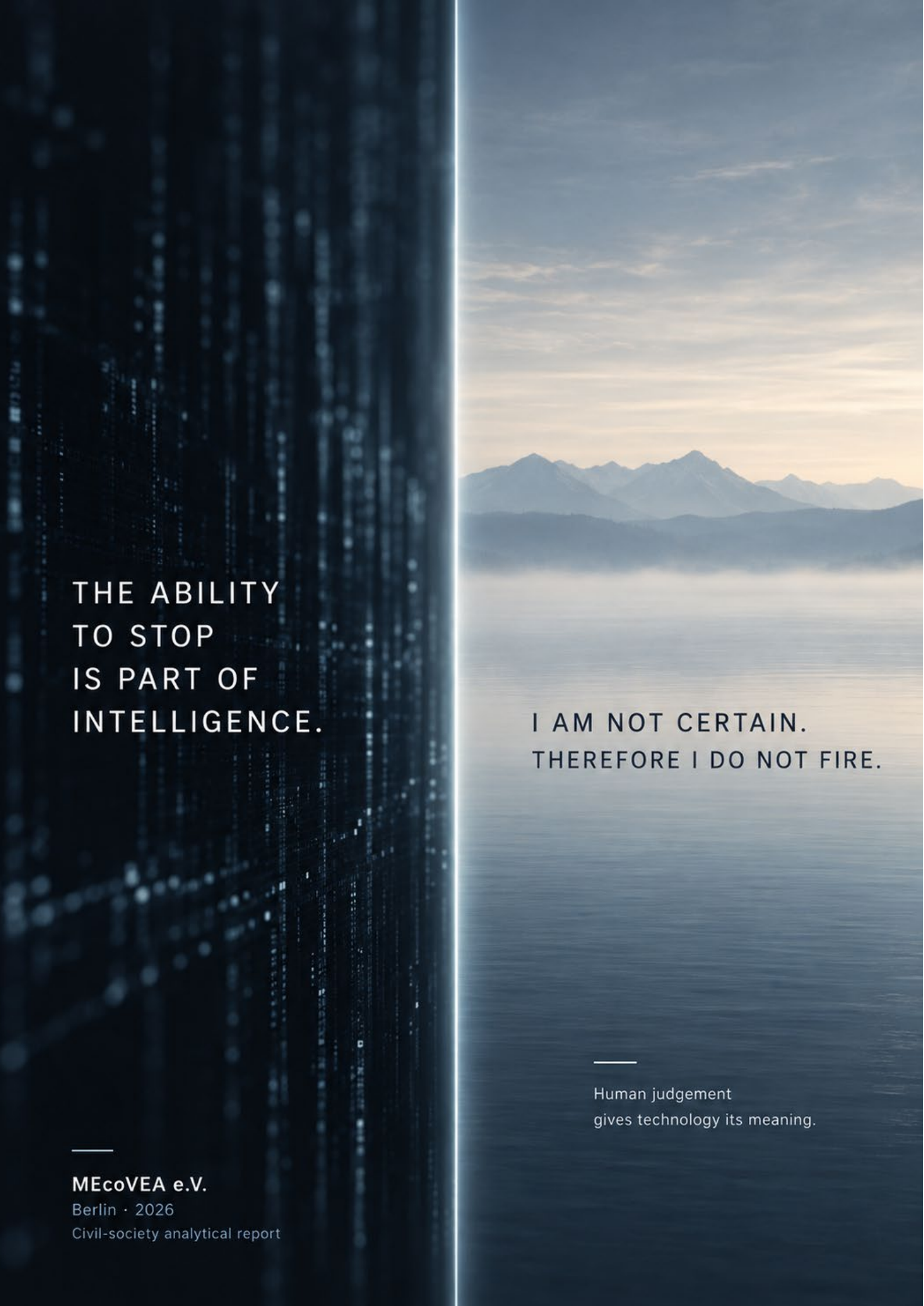
- <https://investors.palantir.com/news-details/2024/Anthropic-and-Palantir-Partner-to-Bring-Claude-AI-Models-to-AWS-for-U.S.-Government-Intelligence-and-Defense-Operations/>
28. **The Wall Street Journal.** U.S. Strikes in Middle East Use Anthropic, Hours After Trump Ban. 28 February 2026.
<https://www.wsj.com/livecoverage/iran-strikes-2026/card/u-s-strikes-in-middle-east-use-anthropic-hours-after-trump-ban-ozN00iClZpfpL7K7ElJ2>
 29. **UN Human Rights Council — Independent International Fact-Finding Mission on Iran.** Iran: Civilians trapped between extreme State repression and violations arising from devastating conflict with US, Israel. 17 September 2026.
<https://www.ohchr.org/en/press-releases/2026/09/iran-civilians-trapped-between-extreme-state-repression-and-violations>
 30. **Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q.** On Calibration of Modern Neural Networks. ICML 2017.
<https://proceedings.mlr.press/v70/guo17a.html>
 31. **Parasuraman, R.; Manzey, D. H.** Complacency and Bias in Human Use of Automation: An Attentional Integration. *Human Factors* 52(3), 381–410, 2010.
<https://journals.sagepub.com/doi/10.1177/0018720810376055>
 32. **Skitka, L. J.; Mosier, K. L.; Burdick, M.** Does automation bias decision-making? *International Journal of Human-Computer Studies* 51(5), 991–1006, 1999.
<https://www.sciencedirect.com/science/article/pii/S1071581999902525>
 33. **Cummings, M. L.** Automation Bias in Intelligent Time Critical Decision Support Systems. AIAA, 2004.
<https://doi.org/10.2514/6.2004-6313>
 34. **UNIDIR Security and Technology Programme.** Artificial Intelligence in the Military Domain and Its Implications for International Peace and Security: An Evidence-Based Road Map for Future Policy Action. 3 July 2025.
<https://unidir.org/publication/artificial-intelligence-in-the-military-domain-and-its-implications-for-international-peace-and-security-an-evidence-based-road-map-for-future-policy-action/>
 35. **OSCE Chairpersonship / Secretariat.** Economic and Environmental Forum focuses on anticipating technologies and science diplomacy for emerging security risks. 10 September 2026.
<https://www.osce.org/chairpersonship/667419>

SOURCE NOTE

Vendor materials are used to establish documented functions and representations, not to prove classified operational use. Media reporting is used only where clearly attributed. Legal propositions rely primarily on treaty text, ICRC materials, UN documents and official national policy documents.

NOTES

Space for observations, questions, references and follow-up.



THE ABILITY
TO STOP
IS PART OF
INTELLIGENCE.

I AM NOT CERTAIN.
THEREFORE I DO NOT FIRE.

Human judgement
gives technology its meaning.

MEcoVEA e.V.

Berlin · 2026

Civil-society analytical report