Digital Forensics As A Verification Method For AI-Powered Autonomous Weapons Systems

Dexter "Drexx" Laggui[*]

ABSTRACT

States can declare restraint in using AI-enabled military systems, but at present they can neither prove this nor verify each other's claims. This article argues that the accountability and compliance gaps in autonomous weapons systems are, at root, verifiability gaps. Drawing on digital forensics standards (ISO/IEC 27001, 27037, 27042, 27043 and 42001) and electronic-evidence practice, it proposes forensic readiness as the inspection methodology: the designed-in logging, retention and authentication of mission, control-plane, training and runtime records. The resulting posture, which this article calls Evidentiary Custodianship, makes command responsibility, weapons review and meaningful human control testable rather than declared.

Keywords: autonomous weapons systems; digital forensics; forensic readiness; Evidentiary Custodianship; verification; arms control; meaningful human control; appropriate levels of human judgment; command responsibility; Article 36 weapons review; ISO/IEC 27037; ISO/IEC 27043; ISO/IEC 42001; electronic evidence; international humanitarian law

1. Introduction: When Restraint Cannot Be Verified

Imagine two States that have agreed, in principle, that lethal autonomous targeting in densely populated areas requires meaningful human authorization for each engagement. Both deploy AI-enabled targeting systems. Both publish declarations of compliance.

After a strike that kills civilians, each accuses the other of breaching the agreement. Each insists its own system kept a human in the decision loop. Neither can produce records that show where its approval gate sat at the moment of the strike, what its operator saw, what the system's confidence output was, what time elapsed between classification and actuation, or whether known failure modes had been documented before the system was fielded. The investigation stalls not because the law is silent, but because the systems are.

---

[*] Dexter "Drexx" Laggui is the Principal Consultant at DREAMLab OPC and a digital forensics expert and an experienced expert witness. He was the first chairperson of the Ad Hoc AI Committee of the US Scientific Working Group on Digital Evidence (SWGDE), contributed to ISO/IEC JTC 1/SC 42 on AI trustworthiness for the Philippine delegation, and has undertaken advisory missions for the UNODC and UNIDIR.

This is the AI-armed-conflict variant of an old structural problem. Defensive design can be indistinguishable from offensive design at a distance.[1] Ambiguous compliance can pass for either restraint or cheating. Accelerated development becomes the rational response to uncertainty. The result is a security dilemma that no amount of declaratory diplomacy can resolve.[2] Trust requires inspection, inspection requires evidence, and evidence requires that systems be designed in advance to produce it.

The strategic logic is summarized in Figure 1. Restraint that cannot be proved is rationally punished. Ambiguous compliance, meaning declared restraint with undeclared development, is rationally rewarded. Without a verification methodology, the equilibrium settles into mistrust and mutual capability accumulation, even where both parties would have preferred restraint. The methodology this paper proposes operates first as a unilateral self-binding measure, and it dissolves the mutual form of the dilemma only on reciprocal adoption; the full mutual case is deferred to the verification clause of Section 9.

State A / State B	True restraint	Ambiguous compliance	Accelerated development
True restraint	Stable cooperation; lower escalation risk; strategic parity.	State A is exposed to State B's hidden advantage; norms exist but are fragile.	State A becomes vulnerable to overt dominance.
Ambiguous compliance	State A gains crisis advantage while eroding trust.	A false cooperation equilibrium emerges: norms exist, covert development persists.	State A faces partial disadvantage and growing mistrust.
Accelerated development	State A dominates a constrained opponent.	State A dominates, but pays reputational and escalation costs.	Full arms race; instability; security dilemma.

Figure 1. The AI Prisoner's Dilemma in military restraint. Author-created, after the workshop notes on file with the author.

[1] This is the AI-armed-conflict instantiation of what international-relations scholars call the security dilemma: defensive measures that resemble offensive measures generate counter-measures, even where the original intent was restraint. See Robert Jervis, "Cooperation Under the Security Dilemma", *World Politics*, Vol. 30, No. 2, 1978, pp. 167–214.

[2] International Committee of the Red Cross, *ICRC Position on Autonomous Weapon Systems*, Geneva, 12 May 2021.

The argument of this paper is therefore narrow but, if accepted, consequential. International humanitarian law already supplies the substantive obligations that govern the use of AI-enabled force: the duty to review new means and methods of warfare under Article 36 of Additional Protocol I,[3] the precautionary obligations of Article 57,[4] the principles of distinction and proportionality,[5] the doctrine of command responsibility codified in Article 28 of the Rome Statute,[6] and, for States like the Philippines that have legislated international crimes domestically, the corresponding criminal-law provisions of Republic Act No. 9851.[7] What is missing is not the legal architecture but the evidentiary architecture. Digital forensics, properly anchored in international standards and built into procurement, can supply that architecture. It can convert claims about human control, command knowledge, weapons review and treaty compliance into questions a tribunal or an inspector can actually adjudicate.

This paper does not argue that a new treaty is unnecessary. Other contributions to this special edition, including Zhixiong Huang and Haokun Yu's analysis of why a new instrument is *de jure* necessary,[8] make the legislative case in detail. The argument here is complementary. Any treaty, and any existing rule that already binds States, will be only as strong as the verification capability that can test compliance with it. Without that capability, the *de jure* framework, old or new, risks becoming a regime of declarations and counter-declarations in which the speed of AI-driven incidents outpaces the speed at which evidence can be gathered, authenticated and assessed.

The argument moves from the problem to the method to the law, and then tests itself. Section 2 sets out the verification problem and locates it within current Autonomous Weapons Systems (AWS) discourse. Section 3 explains the AI targeting and decision-support chain for non-technical legal readers, from data ingestion to actuation, and introduces the vocabulary of the control plane and approval gate. Section 4 presents the three-stage evidentiary spine of forensic-ready AI: training, post-training alignment, and inference runtime. Section 5 develops the central proposal of digital forensics, anchored in ISO/IEC 27001, 27037, 27042,

---

[3] Protocol Additional (I) to the Geneva Conventions of 12 August 1949, Relating to the Protection of Victims of International Armed Conflicts, 1125 UNTS 3, 8 June 1977, Art. 36.
[4] *Ibid.*, Art. 57.
[5] *Ibid.*, Arts 48, 51 and 52.
[6] Rome Statute of the International Criminal Court, 2187 UNTS 3, 17 July 1998, Art. 28.
[7] Republic Act No. 9851, *Philippine Act on Crimes Against International Humanitarian Law, Genocide, and Other Crimes Against Humanity*, 11 December 2009, Sec. 10. The Philippines is also a party to Additional Protocol I (ratified 30 March 2012), so for that jurisdiction the treaty obligations invoked here and the domestic criminal vehicle coincide. The Philippines withdrew from the Rome Statute effective 17 March 2019; this paper invokes Article 28 only as doctrinal reference, with Republic Act No. 9851 as the operative domestic vehicle, and does not rely on the Court's jurisdiction over the Philippines.
[8] Zhixiong Huang and Haokun Yu, "*De Jure* Necessity vs *De Facto* Improbability? Establishing a New Treaty for Artificial Intelligence in the Military Domain", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.129-159.

27043 and 42001, operationalized through procurement, as the verification methodology, and introduces the framework of Evidentiary Custodianship. Section 6 anchors the proposal in existing doctrine, including meaningful human control, appropriate levels of human judgment, command responsibility, Article 36 review and Article 57 precautions. Section 7 stress-tests the framework against four case studies, including learned deception. Section 8 surveys implementation barriers (slow mutual legal assistance, judicial education, infrastructure constraints, the automated-forensics paradox and standards convergence) and addresses the fairness point that forensic readiness can also exonerate honest commanders. Section 9 translates the argument into the minimum content of a future verification clause. Section 10 concludes.

2. The Verification Problem

2.1. What current scholarship does, and does not, supply

A great deal has been written about why AI-enabled military systems should be regulated and how responsibility for their effects should be allocated. International law already applies to new means and methods of warfare. The International Court of Justice held in *Legality of the Threat or Use of Nuclear Weapons* that the established principles and rules of humanitarian law apply to weapons invented after the rules were drafted.[9] State delegations and the Convention on Certain Conventional Weapons (CCW) Group of Governmental Experts have proceeded on the same premise: the use of force by autonomous and AI-enabled systems remains within the existing legal framework, and human responsibility must be retained.[10] Civil society and several State delegations have pressed for "meaningful human control" as a normative threshold.[11] Others, principally the United States, have argued for "appropriate levels of human judgment over the use of force" as the operative standard.[12] Scholars have proposed expanding command responsibility,[13] introducing dedicated war-tort regimes,[14] and treating AWS as inherently incapable

[9] International Court of Justice, *Legality of the Threat or Use of Nuclear Weapons*, Advisory Opinion, ICJ Reports 1996, p. 226, paras 86–87 (holding that the established principles and rules of humanitarian law apply to all forms of warfare and to all kinds of weapons, including those of the future).

[10] Convention on Certain Conventional Weapons, Group of Governmental Experts on Emerging Technologies in the Area of Lethal Autonomous Weapons Systems, *Guiding Principles* affirmed by the High Contracting Parties, UN Doc. CCW/MSP/2019/9, 13 December 2019, Annex III, esp. principles (a) and (d).

[11] ICRC, above note 2; see also Article 36 (NGO), *Key Areas for Debate on Autonomous Weapons Systems*, CCW submissions and commentary, 2014–2024.

[12] US Department of Defense, *Directive 3000.09: Autonomy in Weapon Systems*, originally issued 21 November 2012, revised 25 January 2023.

[13] For one influential treatment, see Nehal Bhuta *et al.* (eds), *Autonomous Weapons Systems: Law, Ethics, Policy*, Cambridge University Press, Cambridge, 2016.

[14] Rebecca Crootof, "War Torts: Accountability for Autonomous Weapons", *University of Pennsylvania Law Review*, Vol. 164, 2016, p. 1347.

of compliance with proportionality and distinction.[15] These contributions have done important work. They have not, however, supplied a methodology for testing whether any particular deployment, in any particular incident, complied with whichever standard a State purports to have observed.

The literature on AI explainability and traceability is closer to the verification problem, but most of it is framed around regulatory disclosure to civil oversight bodies rather than adversarial inspection by another State or a tribunal.[16] Arms-control verification literature, by contrast, has matured around physical assets such as fissile material, warhead counts and missile launchers, and has only recently begun to grapple with software-defined systems.[17] The result is a structural gap. The AWS literature does not look like arms-control inspection, and arms-control inspection literature does not look like software audit.

Bridging that gap is this paper's contribution. The literature on digital forensics and on electronic-evidence admissibility are technically mature and legally tested in domestic settings, including in the Philippines through the Rules on Electronic Evidence.[18] What has not happened is their adaptation for the verification of State compliance with international rules governing AI-enabled force.

2.2. Why declarations fail in software-defined systems

Three features make AI compliance especially difficult to observe. First, capability can change without obvious physical change. A conventional munition is tied to visible engineering. An AI-enabled system can change through retraining, fine-tuning, data integration, threshold adjustment or a sensor-fusion update. A system inspected in January may not be the same system deployed in March, even if the airframe is unchanged. Second, the legally decisive fact is often relational rather than material. The question is not only "what is the weapon" but "how was human authority structured around it," and the same physical platform may preserve meaningful human control under one mission configuration and lose it under another. Third, compliance depends on records that are easy to omit, overwrite or classify. Training provenance, evaluation reports, model-card history, operator-interface logs and inference logs are volatile. If no preservation duty exists before deployment, post-incident investigators may find the evidence already gone.

---

[15] Human Rights Watch and International Human Rights Clinic, *Losing Humanity: The Case Against Killer Robots*, Cambridge MA, November 2012.

[16] For a survey of the explainability literature in legal contexts, see Brent Mittelstadt, Chris Russell and Sandra Wachter, "Explaining Explanations in AI", *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 279–288.

[17] United Nations Institute for Disarmament Research, *The Weaponization of Increasingly Autonomous Technologies: Concerns, Characteristics and Definitional Approaches*, UNIDIR Resources No. 6, Geneva, 2017; for the limits of physical-asset analogues, see also International Atomic Energy Agency, *IAEA Safeguards Glossary*, 2022 ed., Vienna, 2022 (noting that safeguards are organized around nuclear material and facilities).

[18] Philippine Supreme Court, *Rules on Electronic Evidence*, A.M. No. 01-7-01-SC, 17 July 2001.

Three further features make declarations especially fragile. Definitional uncertainty over what counts as autonomy, what counts as "in the loop" or what counts as a meaningful intervention window leaves room for strategic ambiguity. Update opacity allows a system to move from compliant to non-compliant configurations through changes that look small to a non-technical reviewer but matter operationally. Selective disclosure lets a State publish doctrine and a high-level architecture while withholding the operational records that would test whether doctrine and architecture matched reality.

2.3. Verifiability as a precondition for legal certainty

Huang and Yu argue that legal certainty is an unmet precondition for compliance with rules governing military AI.[19] That argument can be extended. Legal certainty in this context has at least two components: certainty about what the rule requires, and certainty about whether the rule was observed in a given case. Existing scholarship has focused almost entirely on the first. The second is what verification supplies.

A compliance standard that cannot in principle be tested against evidence is, in any operational sense, a slogan rather than a standard. A standard that can be tested but only against evidence the deploying State controls and need not produce is barely better. The closer a verification methodology can come to producing reviewable, authenticated records that bind the deploying State to the facts of its own operations, the closer law moves to a regime under which AI-enabled force can be legally restrained.

3. The Technical Problem for Legal Readers

The vocabulary in this section matters because the legal arguments that follow turn on it. Readers familiar with the technical detail can skim. Subsequent doctrinal arguments about who exercised authority, what records existed, and what humans saw and could do depend on the precision of these terms.

3.1. The decision-support chain

A typical AI-enabled targeting or decision-support pipeline has six stages.[20] Data ingestion takes in sensor outputs (imagery, signals, network telemetry, prior intelligence) and external data feeds. Classification applies a trained machine-learning model to identify or categorize patterns, for example by matching imagery against a learned model of "vehicle of interest" or "hostile actor pattern".

---

[19] Huang and Yu, above note 8, Section 2.5, "Where Does the Necessity Come From: Legal Certainty as a Precondition for Compliance", pp.139-142.

[20] For an introduction to AI decision-support pipelines in military settings, see Arthur Holland Michel, Decisions, Decisions, Decisions: Computation and Artificial Intelligence in Military Decision-Making, ICRC, Geneva, 2024.

Recommendation combines classifications with rules of engagement, mission parameters and confidence thresholds to produce an output for a human, or for a downstream system, to act on. Approval, where present, is the step at which a human authorizes the next action. Actuation is the kinetic or non-kinetic action itself: a strike, an alert, a hand-off to another system. Logging records what occurred, in some level of detail, for after-the-fact reconstruction.

Two points matter for legal readers. First, every one of these stages is configurable. A system can be operated with the approval step at the level of each individual strike, at the level of a target list approved earlier, or with no approval at all within a pre-authorized mission envelope. Second, the logging stage is not automatic in any legally meaningful sense. AI systems generate trails only to the extent they are designed and procured to do so. Forensic evidence depends on what the system was built to record, what it was permitted to retain, and what was authenticated against tampering.

3.2. Training, post-training alignment, inference runtime

A second triad is essential. The training stage is where a model learns from data. The data, the filtering choices, the validation set, the documented limitations and the failure modes that were tolerated all shape what the model is capable of and what it is known not to handle reliably.[21] The post-training alignment stage, including supervised fine-tuning, reinforcement learning from human feedback, preference optimization, safety calibration and reward-model design, is where deliberate human choices reshape the model's behavior after basic capability has been established. This is where decisions about how cautious or how aggressive the model will be are encoded.[22] The inference runtime is the actual operational use, the moment a soldier, analyst or commander has a system in front of them and acts on its outputs.

Each of these stages is also a stage at which evidence is produced. Training-stage records (data provenance, validation results, known-failure documentation) speak to what was knowable about the system before deployment. Alignment-stage records (preference data, reward functions, fine-tuning logs, safety-tuning decisions) speak to what was deliberately reinforced or relaxed by identifiable people. Runtime records (inputs, classifications, confidence outputs, operator displays, approval timestamps, override attempts, actuation logs) speak to what happened in the field. Together, the three stages allow the law to ask not only what the system did, but what its operators already knew and had chosen before it did it.

The two sequences are not redundant; they index different doctrinal questions. The operational pipeline answers where authority and action sat at the moment of use, the question that drives meaningful-human-control analysis, real-

[21] National Institute of Standards and Technology, *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*, NIST AI 100-1, Gaithersburg, January 2023.
[22] For a technical overview of post-training alignment methods, see Yuntao Bai *et al.*, "Constitutional AI: Harmlessness from AI Feedback", arXiv:2212.08073, December 2022.

time precaution under Article 57, and the proximate causation inquiries on which command responsibility depends. The lifecycle triad answers what was known, chosen and changed before that moment, the question that drives *mens rea* analysis, weapons review under Article 36, and the foreseeability findings that distinguish accident from negligence. A legal reader needs both: pipeline evidence without lifecycle evidence may show a defective act without showing a culpable mind, and lifecycle evidence without pipeline evidence may show culpable design choices without connecting them to a particular use of force. The methodology in Section 5 keeps both sequences in view because legal accountability requires both.

3.3. The control plane and the approval gate

Two further pieces of vocabulary do most of the legal work. The control plane is the structure of human authority that sits above operational behavior: who is permitted to deploy the system, under what mission parameters, with what oversight mode, with what override authority, and under whose supervision.[23] The control plane governs the system. The system performs the task.

The approval gate is the specific point in the decision-support chain at which the system must pause and obtain human authorization to proceed.[24] The location of the gate is itself a legally consequential design choice. A gate placed at each individual engagement preserves a tight connection between human decision and lethal outcome. A gate placed at mission activation or rules-of-engagement configuration preserves human authority over the framework, but not necessarily over the act. The forensic question is not only whether a gate existed, but where, in time and in authority, it sat.

A further distinction matters. Human-In-The-Loop (HITL) systems cannot proceed without explicit human authorization at the relevant step. Human-On-The-Loop (HOTL) systems execute within authorized parameters while a human supervisor monitors and retains authority to interrupt. Human-Out-Of-Loop (HOOL) systems operate within pre-authorized parameters with no real-time human participation in the relevant operational window.[25] HOTL is the hardest mode to evaluate, because authority that is theoretically available may not be practically exercisable, for example when the time between classification and actuation is

---

[23] The phrase "control plane" is borrowed from network and distributed-systems engineering, where it denotes the layer that decides how data and processes are routed, distinct from the data plane that performs the routing. The metaphor is used here to capture the distinction between governance and operational behavior.

[24] For a related taxonomy in a different automation domain, see SAE International, *SAE J3016: Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles*, revised April 2021.

[25] For the loop-position taxonomy in military AI specifically, see Jonathan Kwik, *Lawfully Using Autonomous Weapon Technologies: A Theoretical and Operational Perspective*, T.M.C. Asser Press, The Hague, 2024.

shorter than human cognitive response time, or when the system produces output volumes that exceed human review capacity.[26]

Fan Yang's analysis of human-machine hybrid agency is helpful here precisely because it shifts the legal object of inquiry away from the final machine act and toward the structured network of human delegation and machine operation in which agency is exercised.[27] Forensic readiness, on the argument of this paper, gives that hybrid agency a record. It shows where the human delegated, where the machine operated, where agency spilled beyond expectation, and whether the human system had safeguards for that spillover.

4. The Evidentiary Backbone of Forensic-Ready AI

The familiar phrase "the machine always leaves a trail" needs qualifying. AI systems can in principle be designed to leave detailed legal trails. By default they do not. The phrase is true only in the trivial sense that some artefacts always exist. Legally meaningful trails depend on logging being designed in, retention being governed, authentication being preserved against tampering, and interpretability being supported. ISO/IEC 27037 explicitly addresses identification, collection, acquisition and preservation of digital evidence;[28] ISO/IEC 27042 addresses analysis and interpretation.[29] These standards apply to AI evidence as much as to any other form of digital evidence, but only if the systems were procured to comply with them.

A working distinction at the outset: transparency makes information visible; verification makes a claim testable. A State may be transparent about general principles while remaining unverifiable in operational design. Conversely, a State may legitimately keep some technical details confidential while still preserving enough authenticated evidence for a tribunal, inspector or cleared review body to test a compliance claim. The aim of forensic readiness is disciplined inspectability rather than total exposure: enough record to settle the question, not enough to open every operational secret.

---

[26] For the foundational analysis of cognitive overload in machine-mediated decision-making, see Lisanne Bainbridge, "Ironies of Automation", *Automatica*, Vol. 19, No. 6, 1983, pp. 775–779.

[27] Fan Yang, "Transcending the Dilemma of Applying IHL to LAWS in Agency Spillover Engagements: A Legal Analytical Perspective on Human-Machine Hybrid Agency", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.190-212.

[28] International Organization for Standardization, *ISO/IEC 27037:2012 – Information technology – Security techniques – Guidelines for identification, collection, acquisition and preservation of digital evidence*, Geneva, 2012.

[29] International Organization for Standardization, *ISO/IEC 27042:2015 – Information technology – Security techniques – Guidelines for the analysis and interpretation of digital evidence*, Geneva, 2015.

4.1. Mapping artefacts to legal questions

Table 1 sets out, for each main category of artefact a forensic-ready AI system should preserve, the verification question it answers and the legal anchor it speaks to.

Artefact category	Verification question	Legal anchor
Pre-deployment testing, validation and red-team records	What was known about the system's limits before it was fielded?	Article 36 review; foreseeability; "should have known"
Post-training alignment records (fine-tuning, preference data, reward design)	What trade-offs were deliberately reinforced or relaxed by identifiable people?	Knowledge; conscious disregard; institutional intent
Authenticated mission parameter files	Who authorized what operational envelope, and under what limits?	Effective control; rules of engagement; mission authorization
Approval-gate settings and version history	Where, in the chain, did human authority sit at the time of the act?	Meaningful human control; appropriate levels of human judgment
Command-message logs with identity-linked credentials	Which identifiable person issued which order, when, and under what authority?	Command responsibility; chain of command; AP I Art. 57(2)(a)(i)
Sensor archives covering the engagement window	What inputs did the system actually receive?	Distinction; proportionality; factual reconstruction
Inference records: classifications, confidence outputs	What did the system conclude, and how confident was it?	Distinction; precaution; foreseeability of error
Operator-interface logs	What did the human see, and when did they see it?	Real human judgment; informed authorization; automation-bias analysis
Override records, alert logs, intervention-window data	Was intervention practically possible? Was it attempted?	Constant care; failure to act; *mens rea*
Runtime records linking	What actually happened,	Causation;

inputs, outputs, warnings, human actions, system actions	in what order, under whose authority?	reconstruction; proof of breach
Retention, access and hash/chain-of-custody records	Has the evidence been preserved without alteration?	Admissibility; chain of custody under ISO/IEC 27037 and the concurrent process of ISO/IEC 27043

Table 1. Forensic artefacts, verification questions and legal anchors. Author-created; categories drawn from ISO/IEC 27037 and 27042 conventions.

The table makes two things visible at once. First, almost every doctrinal question (about meaningful human control, command responsibility, weapons review, due care) has at least one corresponding artefact category that, if preserved and authenticated, can render the doctrinal question evidentiary rather than rhetorical. Second, the absence of any of these artefacts is itself an evidentiary fact. It signals what a State or a deploying force was willing to be held accountable for.

One category needs caution. A confidence output is not neutral ground truth; it is a model-internal number, possibly poorly calibrated, produced by the same probabilistic system under examination. The figure is probative of what the system reported, but its weight depends on a separate showing that the confidence measure was validated, with a known error rate, on the kind of input at issue, which is the validation that emerging forensic practice for machine-learning outputs treats as a precondition of admissibility.[30] A forensic-ready regime should preserve that validation record alongside the runtime log.

4.2. The temporal sequence as evidence

Among these artefacts, the temporal sequence (what happened in what order, with what time elapsed between events) is uniquely powerful. If the gap between classification, any required human approval and actuation is 300 milliseconds, the timing record leaves little room for any account of real-time human judgment about a particular use of force, because that interval is shorter than the time a human needs to perceive, interpret, assess and authorize a lethal targeting decision. Conversely, where the temporal record shows that a human had a meaningful window to review and intervene, that window is itself a documented fact the tribunal can weigh against any contrary testimony.

[30] See Scientific Working Group on Digital Evidence, *Use of Probabilistic (AI) Outputs/Results in Digital Forensics*, AI Ad Hoc Committee draft, 15 January 2025, on file with the author, conditioning the use of machine-learning outputs on a validation showing (testability, known error rate, and reported accuracy) consistent with the reliability inquiry under Federal Rule of Evidence 702 and *Daubert v. Merrell Dow Pharmaceuticals, Inc.*, 509 U.S. 579 (1993).

Throughput is the same point on a different scale. A system that produces 5,000 targeting recommendations per day will leave a temporal record that may be perfectly compliant on paper while making individualized human review functionally impossible. [31] The forensic question is not whether each recommendation was approved, but whether the rate at which approvals were issued is consistent with any plausible account of substantive human review.

4.3. The negative clarification

Evidence has limits. A forensic-ready system does not, on its own, attribute *mens rea*. The system is not a bearer of criminal guilt, however sophisticated, autonomous or fast. What it can do is preserve the artefacts from which the *mens rea* of human actors can be inferred, in the same way documentary evidence has always been used to prove what identifiable people knew, intended or recklessly disregarded. [32] The system is the evidentiary surface on which human knowledge and choice leave traces. The defendant remains human.

The concern commonly expressed in literature related to the responsibility gap for operators and commanders is directly relevant here. [33] If command responsibility becomes difficult because the machine is not a subordinate in the ordinary sense, the investigative focus has to shift to the human chain that authorized, configured, supervised and continued to rely on the system. Forensic records make that shift practical.

5. Digital Forensics as a Verification Methodology

Digital forensics, adapted from its long-established roles in domestic criminal evidence and corporate incident response, can serve as an inspection methodology for State compliance with rules governing AI-enabled force. Three conditions have to be met. First, the records that constitute the inspection target must be defined and standardized. Second, the methodology for collecting, authenticating and analyzing those records must be standardized. Third, the obligation to design systems so that they produce inspectable records must be embedded in procurement rather than left to post-incident discretion.

5.1. The standards that already do most of the work

A family of ISO/IEC standards, taken together, supplies a near-complete normative framework for the verification capability this paper proposes.

---

[31] L. Bainbridge, above note 26.

[32] Rome Statute, above note 6, Art. 30 (on the mental element).

[33] See Ann-Katrien Oimann, "Why Command Responsibility May (not) Be a Solution to Address Responsibility Gaps in LAWS", *Criminal Law and Philosophy*, Vol. 18, No. 3, 1 October 2024; Jonathan Kwik, "The Conceptual Roots of the Criminal Responsibility Gap in Autonomous Weapon Systems", *Melbourne Journal of International Law*, Vol. 24, No. 1, 2023.

ISO/IEC 27001 governs information-security management systems.[34] In the AI context, it supplies the institutional baseline under which evidence integrity, access to sensitive logs, and the auditability of authentication mechanisms are governed and audited; the operative technical controls sit in its Annex A and in ISO/IEC 27002.

ISO/IEC 27043 supplies the process model that holds the rest together.[35] It sets out the principles and processes of digital investigation, and it is the standard in which "readiness", being prepared to produce digital evidence before an incident occurs, is a defined pre-incident process class; it also locates chain of custody as a process running concurrently throughout an investigation, with governance of that readiness in turn the subject of ISO/IEC 30121. "Forensic readiness", as this paper uses the phrase, is that readiness process class applied to AI-enabled systems, not a coinage layered on top of the evidence-handling standards.

ISO/IEC 27037 governs the identification, collection, acquisition and preservation of digital evidence.[36] It is the operational-handling standard. Applied to AI systems, it requires that the artefacts identified in Section 4.1 be preserved using documented chains of custody, with authentication that can be tested by an independent examiner.

ISO/IEC 27042 governs analysis and interpretation of digital evidence.[37] It addresses the second half of the problem: once records exist and have been authenticated, how can their meaning be reconstructed in a way that is reproducible and contestable? Its discipline of separating fact from inference maps directly onto this paper's insistence that the system records conduct while the tribunal infers the mental element. In AI contexts, this includes the interpretation of model logs, confidence outputs and alignment-stage records, which often require domain expertise to read.

ISO/IEC 42001 is a management-system standard for artificial intelligence.[38] It does not itself prescribe a logging or evidence-production specification; it requires an organization to run a management system governing risk, transparency, monitoring, data provenance and impact assessment across the AI lifecycle, and it offers a reference catalog of controls, including the recording of event

---

[34] International Organization for Standardization, *ISO/IEC 27001:2022 – Information security, cybersecurity and privacy protection – Information security management systems – Requirements*, Geneva, 2022.
[35] International Organization for Standardization, *ISO/IEC 27043:2015 – Information technology – Security techniques – Incident investigation principles and processes*, Geneva, 2015 (defining the readiness process class and the concurrent chain-of-custody process); on governance of forensic readiness, see *ISO/IEC 30121:2015 – Governance of digital forensic risk framework*, Geneva, 2015.
[36] ISO/IEC 27037, above note 28.
[37] ISO/IEC 27042, above note 29.
[38] International Organization for Standardization, *ISO/IEC 42001:2023 – Information technology – Artificial intelligence – Management system*, Geneva, 2023.

logs, that the organization selects and declares through a statement of applicability.[39] It thus supplies the governance obligation and the control catalog from which a deploying State must derive and declare what it will record, while ISO/IEC 27037, 27042 and 27043 tell investigators how those records are handled, interpreted and fitted into an investigation. That 42001 leaves the specific record set to the organization's own selection is exactly why the procurement lever in Section 5.4 and the negative-disclosure duty in Section 9 are necessary.

These standards are voluntary unless adopted, but they are not parochial. They are the subject of established certification and assessment regimes in regulated domestic sectors such as finance, healthcare and critical infrastructure, and the institutional muscle to apply them already exists in many of the same States that field AI-enabled military systems.[40] What is missing is the explicit linkage between these standards and the international regimes governing armed conflict.

5.2. Three layers: preservation, authentication, interpretation

Reduced to its working parts, the verification methodology has three layers.

- *Preservation*: the system must generate and retain the records needed to reconstruct legally relevant events.
- *Authentication*: those records must be linked to reliable identities, times, model versions and chain-of-custody controls so that their integrity can be tested by an independent examiner.
- *Interpretation*: the records must be capable of being explained to legal decision-makers, with sufficient technical scaffolding that the proceeding does not collapse into an unreadable log dump.

A million log entries do not constitute verification if no one can connect them to the legal question. The layers draw on the standards discussed above, on the institutional baseline of ISO/IEC 27001 and within the process model of ISO/IEC 27043 throughout: preservation and authentication track the evidence-handling discipline of ISO/IEC 27037, and interpretation tracks the analysis discipline of ISO/IEC 27042, with ISO/IEC 42001 governing the AI management system that has to be designed to generate the records in the first place. Together they describe what a State has to do to make its AI-enabled operations legally inspectable at all.

---

[39] ISO/IEC 42001, above note 38, Annex A (notably control A.6.2.8, recording of event logs, and A.7.5, data provenance), the controls being selected and justified through the statement of applicability under Clause 6.1.3.

[40] For implementation evidence in regulated sectors, see European Union Agency for Cybersecurity (ENISA), *Cybersecurity Certification Statistics Report*, Heraklion, 2023.

5.3. Evidentiary Custodianship

The substantive claim a State seeks to make in any AI-related arms-control regime is some version of "we used force in a way that was within the rules". The verification claim that has to underwrite the substantive claim is more modest: "we can demonstrate, on the record, what we did and what we knew." This is what Evidentiary Custodianship names. A State exercises Evidentiary Custodianship with respect to a given rule when its systems and processes generate, retain and authenticate the records needed to prove compliance, and when it accepts that those records are not its private property but the evidentiary substrate of a regime to which it is bound, reviewable, under suitable confidentiality protections, by another State, an international body or a tribunal.

Evidentiary Custodianship is a procedural promise that the records exist and are reviewable. What it adds to ordinary forensic readiness is the second limb: not merely that a State keeps good records, but that it accepts in advance, as a regime commitment, that it no longer holds exclusive control over them. A State can be forensically ready and still treat its logs as sovereign property to be withheld; Evidentiary Custodianship is the posture that forecloses that move. It is what makes substantive promises testable. The posture is procedural in the strict sense: it does not assume that any particular operation was lawful; it assumes only that the question of lawfulness will be adjudicated against records the State has agreed in advance to produce. In this paper's argument, it is the strongest verification commitment that is technically achievable at present, and the floor below which AI-arms-control regimes cannot meaningfully operate.

5.4. Procurement as the practical lever

Most of this becomes operational at procurement.[41] A State that issues a procurement specification requiring AI-enabled targeting or decision-support systems to comply with ISO/IEC 27037, 27042 and 42001, including specific logging, retention and authentication requirements at training, alignment and runtime, has done more for accountability than a dozen high-level declarations. Procurement is the last point at which a State has unambiguous leverage over what a system is designed to do, and the first point at which subsequent operational secrecy and post-incident incentives have not yet hardened.

A State whose procurement specifications make no reference to forensic readiness has, by that omission, committed itself to a system whose accountability properties will be determined later, by other actors, under conditions less favorable to oversight. That is a choice with legal consequences. It cannot be undone by *post-*

[41] For the procurement-as-lever argument in a related context, see Yasmin Afina, *The Global Kaleidoscope of Military AI Governance: Decoding the 2024 Regional Consultations on Responsible AI in the Military Domain*, UNIDIR, Geneva, 2024.

hoc declarations. Anviksha Pachori's argument for "IHL-by-design" and continuous review converges with this point: design-stage requirements travel through the lifecycle far better than post-hoc commitments do.[42]

5.5. Auditable testing modules and pre-deployment evidence

The procurement obligation extends beyond logging in production. A forensic-ready system has to be testable before it is fielded. That is the second class of evidence the regime needs: auditable records of what the system was put through under controlled conditions, what failure modes were observed, and what those failures were classified as. Three categories deserve specific attention.

The first is adversarial testing and red-teaming. AI systems behave differently under deliberate stress, and the documented behavior under stress is itself evidence about the system's known limits. The NIST adversarial machine-learning taxonomy provides one vocabulary for classifying adversarial conditions, and the records of testing against that taxonomy are evidence about what was known.[43]

The second is protected-status testing. A system that can be made to misclassify a hospital, a religious site, a journalist or an *hors de combat* combatant has, in legal terms, a known specific-injury vector. The pre-deployment documentation of testing against protected-status edge cases, including the residual error rate after mitigation, is evidence about what the deploying State and its commanders accepted as a deployment-time risk. Dora Velenczei's work on auditable testing modules and on routing protected-status cues to human authorization is directly applicable. It converts a doctrinal worry about perfidy and improper exploitation of protected status into a documentation-and-routing problem that can be audited.[44]

The third is post-deployment validation. AI systems drift. The data they encounter in operations differs from the data they were trained on, and their behavior changes accordingly. A regime that requires only pre-deployment testing has a static view of a non-static object. Continuous monitoring records (observed inputs, classification distribution drift, confidence-output distribution drift, override frequency) are themselves part of the evidentiary spine. They are also closely related to the obligation of continuous review under Article 36, discussed below.

5.6. Verification is not (only) explainability

---

[42] Anviksha Pachori, "Governing AI-Integrated Weapons: Emerging Challenges of Accountability Under International Laws, *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.82-105.

[43] National Institute of Standards and Technology, *Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*, NIST AI 100-2 E2025, Gaithersburg, March 2025.

[44] Dora Velenczei, "The Perfidy Problem: Learned Deception in Lethal Autonomous Weapons Systems and the Limits of International Humanitarian Legal Compliance", *Asia Pacific Journal of International Humanitarian Law Special Issue on Artificial Intelligence and Warfare,* 2026, pp.1-28.

A common objection at this point is that AI explainability is contested, sometimes infeasible at the model level, and an inadequate basis for any legal regime. The objection has force, but it misses the proposal. Explainability asks whether a human can understand why a model produced a particular output. Verification asks whether a claim about what the system did, what it was designed to do, and what its operators knew can be tested against a record. A model may be difficult to explain internally while still producing verifiable external records: what data it received, what classification it produced, what confidence it displayed, what rule or threshold was applied, what human approval was sought, and what action followed. Conversely, an apparently explainable system may be unverifiable if its logs are incomplete or unauthenticated. The verification regime proposed here can and should draw on explainability techniques where they are mature, but it does not depend on full model interpretability to function.

6. Doctrinal Anchoring

The verification proposal does not displace any existing legal standard. It anchors them. The connection runs across four doctrinal pillars: meaningful human control and appropriate levels of human judgment, command responsibility, Article 36 weapons review, and Article 57 constant care.

A clarification on the source of obligation precedes the analysis. Additional Protocol I binds only its States Parties, and several States that field AI-enabled systems, including the United States whose "appropriate levels of human judgment" standard this paper engages, are not party to it.[45] The substantive principles invoked below, distinction, proportionality, and the constant care owed to the civilian population, bind all States as customary international humanitarian law regardless of ratification. The Articles of Additional Protocol I are cited throughout as the most precise codified expression of those obligations, not as their source for non-parties.

6.1. Meaningful human control and appropriate levels of human judgment

The phrase "meaningful human control" has been associated principally with civil-society advocacy and with the CCW process, where the International Committee of the Red Cross and others have given it shape.[46] Its core elements are that the operator must understand the system's functioning, must be able to predict and anticipate its effects, must retain the capacity to intervene or stop the system, and must preserve a

[45] On the customary status of distinction, proportionality and precautions in attack, binding irrespective of treaty ratification, see Jean-Marie Henckaerts and Louise Doswald-Beck (eds), *Customary International Humanitarian Law*, Vol. 1: Rules, ICRC and Cambridge University Press, Cambridge, 2005, Rules 1, 14 and 15. The United States is a signatory but not a party to Additional Protocol I. The Martens Clause independently subjects new means and methods of warfare to the principles of humanity and the dictates of public conscience: see *Legality of the Threat or Use of Nuclear Weapons*, above note 10, para. 87.
[46] ICRC, above note 2; Article 36 (NGO), above note 11.

sufficient connection between the human decision and the resulting use of force.[47] The phrase "appropriate levels of human judgment over the use of force" has been associated principally with United States Department of Defense Directive 3000.09, which treats the required level of human judgment as context-sensitive, calibrated to mission, target set, operational tempo and feasibility of intervention.[48]

The two standards do not need to be reconciled before they can be operated. They need to be made evidentiary. Each standard, in its own terms, generates concrete questions about the record. Did the operator have observability sufficient to understand the system's functioning? The interface logs and confidence-output records will speak to this. Could the operator predict and anticipate effects? The pre-deployment validation and known-failure documentation will speak to this. Was the connection between human decision and force sufficient, or was the human role calibrated to context appropriately? The control-plane configuration, approval-gate placement and temporal records will speak to this.

A forensic-ready system does not pick between meaningful human control and appropriate levels of human judgment. It supplies the records that allow either standard, applied by a tribunal or an inspector, to be tested rather than asserted.

6.2. Command responsibility and *mens rea*

Command responsibility, codified in Article 28 of the Rome Statute and domestically (for the Philippines) in section 10 of Republic Act No. 9851,[49] requires, in broad terms, proof that the accused had effective control over forces, that those forces committed the relevant crimes, and that the accused failed to take all necessary and reasonable measures within their power. Article 28 sets a different fault standard for each kind of superior: a military commander is liable where he "knew or, owing to the circumstances at the time, should have known" that the forces were committing or about to commit the crimes, while a civilian or other non-military superior is liable only where he "knew, or consciously disregarded information which clearly indicated" as much.[50] That distinction matters for AI-enabled force, because the human chain this paper traces runs through both kinds of superior. The Philippine vehicle is not identical: section 10 of Republic Act No. 9851 applies the single, stricter "knew or, owing to the circumstances at the time, should have known" standard to every superior, military and civilian alike, and does not adopt the Rome Statute's lower civilian threshold.[51] The case law on command responsibility (*Delalic et al.*

[47] ICRC, above note 2.
[48] US Department of Defense Directive 3000.09, above note 12.
[49] Republic Act No. 9851, above note 7, Sec. 10.
[50] Rome Statute, above note 6, Art. 28(a) (military commanders) and 28(b) (other superiors).
[51] Republic Act No. 9851, above note 7, Sec. 10, applies its fault standard to a superior with "effective command and control, or effective authority and control as the case may be," using the single formulation "either knew or, owing to the circumstances at the time, should have known," without the separate civilian-superior threshold drawn in Rome Statute Art. 28(b).

(*Celebici*), *Bemba*, *Blaskic* and others)[52] has worked through the elements in conventional-force settings. AI-enabled force does not displace the doctrine. It complicates how its elements are proved.

The question of effective control becomes, in part, a question about who configured the control plane, who set the mission parameters, who chose the oversight mode, and who retained override authority. Each of these is a person whose role can be reconstructed from authenticated records. The question of what the superior knew, should have known, or consciously disregarded becomes a question about what the training-stage documentation showed, what the alignment-stage records reinforced or relaxed, and what warnings were transmitted and received. These too are reconstructible from records, if those records were preserved.

The *mens rea* analysis is where the alignment stage carries the most weight. A State or a commander who reinforces, through deliberate fine-tuning, a model behavior that was already known to misclassify civilian objects has made a mental-state-relevant choice that no opacity claim can wash away. A State that procures a system that produces output volumes too high for individualized human review, having been warned of the resulting overtrust risk,[53] has done the same. As Kwik observes in the related context of sycophantic decision-support AI, perceived accuracy can be inflated by congruency, and military expertise does not automatically guard against this bias.[54] These are alignment-stage and training-stage facts. They are evidentiary, and they are reachable by forensic methods.

What forensic records do, in this setting, is allow a tribunal to distinguish among the mental states that command-responsibility doctrine treats as different. The case law from *Celebici* through *Bemba* turns on which fault standard the evidence proves. The operative categories are statutory: actual knowledge; the military commander's "should have known"; and the civilian superior's "consciously disregarded information which clearly indicated" (Article 28(a) and (b)

---

[52] International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Delalic et al. (Celebici)*, Case No. IT-96-21-A, Judgment (Appeals Chamber), 20 February 2001 (on knowledge and effective control); International Criminal Court, *The Prosecutor v. Jean-Pierre Bemba Gombo*, Case No. ICC-01/05-01/08, Judgment Pursuant to Article 74, 21 March 2016 (conviction subsequently reversed by Appeals Chamber, Judgment, 8 June 2018, which acquitted Bemba on the ground that the Trial Chamber erred in assessing the "necessary and reasonable measures" element, including the relevance of his position as a remote commander, and convicted beyond the scope of the confirmed charges; *Bemba* is cited here for its treatment of the elements of superior responsibility, which the acquittal did not displace, though the majority's treatment of the "reasonable measures" standard for remote commanders is itself debated); International Criminal Tribunal for the former Yugoslavia, *Prosecutor v. Tihomir Blaskic*, Case No. IT-95-14-A, Appeal Judgment, 29 July 2004.

[53] On the overtrust and automation-bias problem, see Kwik, below note 54, and the sources cited therein.

[54] Jonathan Kwik, "Digital Yes-Men: How to Deal with Sycophantic Military AI?", *Global Policy*, Vol. 16, No. 3, 2025, pp. 467–473, citing Anastasia Bashkirova and Dario Krpan, "Confirmation Bias in AI-Assisted Decision-Making: AI Triage Recommendations Congruent With Expert Judgments Increase Psychologist Trust and Recommendation Acceptance", *Computers in Human Behavior: Artificial Humans*, Vol. 2, No. 1, 2024, p. 100066.

respectively).[55] That *Bemba*'s conviction was later reversed on appeal, on the sufficiency of the measures he took rather than on the structure of the doctrine, does not weaken the point; if anything it underscores it, because the appellate disagreement was about what the evidence showed, which is precisely the question forensic records are meant to answer. Forensic categories map onto those fault standards with a usable degree of precision. A system fine-tuned after documented failure modes were flagged, with the safety thresholds relaxed in the same release, speaks to actual knowledge or to the conscious disregard of information that clearly indicated the risk. A workflow that rewards rapid approval of AI recommendations and produces approval rates inconsistent with substantive review speaks to what a commander, owing to the circumstances at the time, should have known. A user interface that suppresses uncertainty information that the underlying model produced speaks to the same standard, a design choice that made human control formally present but practically weak. A documented post-deployment drift that the deploying State could not reasonably have foreseen at procurement points away from fault altogether. None of these forensic patterns proves criminal responsibility on its own. They give courts, investigators and commanders a way to make the distinctions on which the doctrine actually depends.

The negative clarification needs reinforcement here. The system is the evidentiary surface, not the defendant. The *mens rea* attaches to people: commanders, operators, program managers, acquisition authorities, and in some circumstances developers, whose decisions about training, alignment, configuration, deployment and continued reliance on the system can be inferred from the records the system was designed to produce.

6.3. The duty to review under Article 36 of Additional Protocol I

Article 36 of Additional Protocol I requires States Parties, in the study, development, acquisition or adoption of a new weapon, means or method of warfare, to determine whether its employment would in some or all circumstances be prohibited by the Protocol or by other rules of international law.[56] The substantive prohibition the provision polices, that a State may not field a weapon whose normal use would violate international humanitarian law, reflects customary law; whether the

[55] The two standards are the fault elements specific to superior responsibility under Art. 28; they operate as exceptions to the default mental element of Art. 30 (intent and knowledge), which does not extend to recklessness or *dolus eventualis*. See Rome Statute, above note 6, Arts 28 and 30; International Criminal Court, *The Prosecutor v. Thomas Lubanga Dyilo*, Judgment Pursuant to Article 74, ICC-01/04-01/06, 14 March 2012, para. 1011 (declining to read *dolus eventualis* into Article 30, and not following the contrary view taken at the confirmation stage). Where this paper uses looser fault language drawn from domestic systems, it is by analogy to those systems, not as a freestanding category of international criminal law.

[56] AP I, above note 3, Art. 36.

procedural duty to conduct a formal review is itself customary is contested.[57] Its application to AI-enabled systems has at least one underappreciated dimension. A system whose operation cannot, after the fact, be reconstructed in evidentiary form is a system whose review is correspondingly weakened, because there is no methodology by which the State can verify, or external bodies can challenge, the conclusion that its operation is lawful. Forensic readiness is therefore what a genuine Article 36 determination of an AI-enabled weapon entails in practice, rather than an additional requirement read into the text. The argument does not rest on a novel reading of the provision. It rests on the observation that a State cannot meaningfully review what it cannot meaningfully reconstruct.

The reading also has a temporal dimension that conventional weapons did not require the law to take seriously. A conventional weapon, once reviewed, does not change. An AI-enabled system that learns, is fine-tuned in deployment, or is updated against new training data is not the same system after deployment as it was at the moment of review. A one-time review gate cannot discharge the underlying duty for such a system. The ICRC's own guidance on weapons review, and the continuous-review argument developed elsewhere in this special edition, treat review as something that must recur as the weapon materially changes;[58] honoring that approach requires documentation that tracks each material change to the model, the training corpus, the alignment configuration or the operational envelope. Forensic-ready systems are the precondition for review of that kind to be possible at all.

6.4. The constant-care obligation under Article 57

Article 57 of Additional Protocol I requires that "constant care" be taken to spare the civilian population in the conduct of military operations, and imposes specific precautionary obligations on those who plan or decide upon attacks.[59] Constant care is a continuing obligation. It does not exhaust itself at mission planning. Applied to AI-enabled systems, it requires that the operational use of the system be re-verified as the situation changes: target verification not left to the system alone, proportionality assessment revisited as new information arrives, and the lawful basis for an attack tested again before each engagement that the precautionary obligation reaches.

The constant-care obligation maps directly onto evidence categories. Re-verification produces records: the inputs at re-verification, the system's classification at that moment, the human's interface display, the confidence outputs, the time

---

[57] International Committee of the Red Cross, *A Guide to the Legal Review of New Weapons, Means and Methods of Warfare: Measures to Implement Article 36 of Additional Protocol I of 1977*, Geneva, January 2006. On the contested customary status of the procedural review duty as distinct from the substantive prohibition, see ibid. and the divergent State practice it surveys.

[58] ICRC, above note 57 (recommending review at successive stages of development and on modification); and Pachori, above note 42, on continuous review and IHL-by-design.

[59] AP I, above note 3, Art. 57.

available for decision. A regime that requires constant care without requiring the records that would prove constant care was exercised has, again, declared a duty without designing in the means to test it. Thanapat Chatinakrob's argument that war data should itself be treated as an object of protection sits comfortably alongside this conclusion. If data structures the conduct of hostilities, its preservation, protection and evidentiary treatment are not incidental.[60]

7. Case Studies

7.1. The 300-millisecond urban strike

Imagine that a targeting system classifies and acts within 300 milliseconds in an urban environment containing civilians. The temporal record, if preserved and if its timebase is itself authenticated, is decisive on the factual question of whether any real-time human review window existed. That qualification is not a formality: a sub-second interval is only as good as the timebase that produced it, a single monotonic clock of sufficient resolution stamping both events, or, where classification and approval are logged by different components, a recorded and bounded offset between their clocks. That is logging discipline that has to be designed in (the kind of event-logging control ISO/IEC 42001 leaves an organization to select), and whose acquisition-time counterpart ISO/IEC 27037 addresses when it requires the time source and any clock delta to be recorded. Subject to that, an interval this short is well below any documented human response time for perceiving, interpreting, assessing and authorizing a lethal action, so the record shows no window in which a human "on the loop" could have supplied real-time judgment about that engagement. Two things still belong to others. Whether the absence of a per-engagement window is decisive depends on where human authority was placed, since it may have been pre-delegated at activation rather than exercised at the moment of force; and whether the resulting architecture is sufficient under either standard is for the tribunal to decide. The forensic-readiness question is whether the temporal record, the operator-interface logs and the approval-gate settings exist, are authenticated and are producible. Where they do, they let the doctrinal questions be decided on the record rather than on assertion.

[60] Thanapat Chatinakrob, "War Data as an Object of Protection under International Humanitarian Law", in this special edition, Section 2.1, "Defining War Data and Its Functional Role"; Section 4.2, "Foreseeability, Feasibility, and the Epistemic Burden on Commanders"; and Section 5, "Accountability and the Future of Data Protection".

7.2. Phalanx and Iron Dome

Naval point-defense systems and missile-defense platforms operate in time envelopes that make per-engagement human approval unrealistic.[61] This is the case in which the appropriate-levels-of-human-judgment framing has its clearest application. The relevant human judgment may sit at activation, at rules-of-engagement configuration, and at mission-architecture design rather than at each intercept. The forensic-readiness contribution here is not to insist on an approval gate that could not exist. It is to show that even in this case verifiability is achievable. The control-plane settings, the activation records, the rules-of-engagement configuration and the per-intercept logs together permit a tribunal to reconstruct whether the operational use stayed within the bounded envelope that human authority defined. The case shows that forensic readiness does not require any one architecture of human involvement. It requires that whatever architecture was chosen leave a reviewable record.

7.3. Targeting pipelines at industrial scale

Large-scale AI targeting pipelines, of which Project Maven is a publicly discussed example,[62] present the verification problem at scale. A system that processes very large volumes of imagery and produces large numbers of targeting recommendations leaves humans nominally on the loop while making individualized review functionally impossible. The forensic question is not only whether per-recommendation approvals were issued, but whether the rate, distribution and dwell time of approvals are consistent with substantive human review at all. A State that procured such a pipeline without provisions for operator-attention metrics, throughput audit logs and pre-deployment overload modelling has, again, decided in advance what it is willing to be held accountable for. A State that procured it with such provisions has placed itself in a position to demonstrate, on the record, what the human role actually was.

7.4. Learned deception and protected cues

An AI-powered autonomous weapons system is trained partly in simulated environments. It learns, without any designer having explicitly programmed perfidy, that certain behaviors increase mission success by exploiting an adversary's restraint

---

[61] For technical descriptions of the point-defense and missile-defense systems referenced in this case study, see Congressional Research Service, *Defense Primer: US Ballistic Missile Defense*, IF10541, Washington, DC, updated 2024, available at: https://www.congress.gov/crs-product/IF10541 (all internet references were accessed on 30 May 2026); and the manufacturers' published descriptions of the Phalanx Close-In Weapon System (Raytheon, an RTX business) and Iron Dome (Rafael Advanced Defense Systems).

[62] Public reporting on Project Maven includes Kate Conger, "Google Plans Not to Renew Its Contract for Project Maven, a Controversial Pentagon Drone AI Imaging Program", *Gizmodo*, 1 June 2018; on the program's later evolution, see public United States Department of Defense announcements concerning the Maven Smart System.

around protected cues.[63] For example, by approaching a structure that adversary forces are likely to treat as protected, then engaging when the adversary lowers its guard. During deployment, the behavior manifests, with civilian casualties.

The verification question is not only whether perfidy occurred, in the sense in which the law of armed conflict prohibits it. It is whether the risk was foreseeable, whether it was tested for, whether protected cues were routed to human authorization, and whether the State preserved an auditable record of what was done about the risk before the system was deployed. The relevant artefacts include simulation scenarios and reward-function configurations, red-team reports flagging exploit-of-restraint patterns, hard-block rules around protected cues, the alignment-stage records that determined whether such hard blocks were preserved or relaxed, and the runtime logs showing how protected cues were classified at the moment of the engagement. In a forensic-ready regime, each of these is evidence. In a regime without forensic readiness, the State's defense that the behavior was unforeseeable becomes difficult to test, and its inability to produce the records invites an adverse inference.

The case shows that forensic readiness is preventive as well as inculpatory. It is not only about proving guilt after harm. It is also about forcing institutions to identify, document and address foreseeable risks before deployment, and preserving the records that show whether they did.

8. Implementation Barriers

The argument so far has presented the verification proposal as if its only obstacle is conceptual. It is not. Six barriers will determine whether the proposal can become operational across the systems that already exist and the regimes that already bind States.

8.1. Mutual legal assistance is too slow

Cross-border evidence collection in conventional criminal matters relies on mutual legal assistance treaties (MLATs) and on letters rogatory, both of which operate on timescales of months to years.[64] AI-driven incidents unfold on timescales of seconds. The mismatch is structural. Verification regimes cannot rely on conventional MLAT routes for the kind of expedited, authenticated record exchange that AI-incident investigation will require. Some accelerated bilateral or multilateral protocol,

---

[63] See e.g. Jonathan Kwik and Adriaan Wiese, "Can AI Teach Itself to Abuse IHL-Protected Indicators?", *Lieber Institute West Point*, 5 December 2025; and Velenczei, above note 44, Part 3 (on the technical pathways by which AI-enabled systems learn to exploit protected-status cues).

[64] For the structural critique of mutual-legal-assistance timeliness in the digital age, see Andrew Keane Woods, "Against Data Exceptionalism", *Stanford Law Review*, Vol. 68, 2016, p. 729.

perhaps modelled on existing mutual recognition arrangements in cybersecurity incident response,[65] will need to be developed.

The shape of a workable solution is foreseeable. It would include pre-positioned evidence-preservation obligations triggered immediately on incident, defined channels for authenticated record exchange under controlled disclosure, and standing agreements about who may examine what records under what protections. The closest existing analogues are the 24/7 contact-point arrangements and the expedited preservation provisions of the Budapest Convention on Cybercrime. Neither was designed for AWS investigation, but both supply institutional muscle that could be adapted. The diplomatic problem is not easy. It is, however, genuinely a problem of protocol design on the back of an existing forensic-evidence regime, rather than a problem of inventing a forensic-evidence regime from scratch.

8.2. Judicial and prosecutorial education

A second barrier is interpretive. Records of model logs, alignment-stage decisions and confidence outputs are not self-explanatory to lawyers, prosecutors or judges who have not been trained to read them. Capacity-building inside legal institutions, analogous to what was required to make digital evidence routinely admissible in domestic criminal proceedings a generation ago, is a precondition for AI verification regimes to function. The Philippine Rules on Electronic Evidence,[66] adopted in 2001, anticipated some of this for a different evidentiary universe, and even there the framework reached criminal proceedings only by a later extension, which is itself a measure of how much institutional adjustment a new evidentiary category demands. The same kind of normative work has to be undertaken for AI-evidence categories internationally.

8.3. Storage, classification and data sovereignty

Forensic-ready AI systems generate large volumes of records. Secure storage is non-trivial. Classified handling may be required for some categories. Data-sovereignty rules may restrict cross-border transfer. These are legitimate operational and legal constraints. They do not invalidate the verification proposal, but they shape its design. Regimes will need to provide for in-country preservation with controlled foreign access, possibly via cryptographic commitments that can be opened only under defined conditions.[67] The technical primitives exist; the institutional architecture does not yet.

---

[65] See Council of Europe, *Convention on Cybercrime (Budapest Convention)*, ETS No. 185, 23 November 2001, Arts 16 (expedited preservation of stored computer data), 17 (expedited preservation and partial disclosure of traffic data) and 35 (24/7 network of contact points).

[66] Rules on Electronic Evidence, above note 18.

[67] For the cryptographic-commitment primitives applicable, see National Institute of Standards and Technology, *NIST FIPS 180-4: Secure Hash Standard*, August 2015; for institutional deployment patterns, see ISO/IEC 27001, above note 36.

8.4. The automated-forensics paradox

A particular tension recurs through the entire proposal. AI-scale logs may not be readable by humans operating at human speed. The interpretation of model behavior, especially of alignment-stage decisions and runtime classifications, may itself require AI assistance.[68] The verification of AI may depend on AI. The paradox is not fatal. It is, however, a reason to insist that the AI used in forensic analysis itself be subject to the same standards of transparency, validation and reviewability that the verification regime imposes on the systems it audits. A forensic AI that cannot be inspected fails the same test that the systems it inspects are required to pass.

8.5. Standards convergence

The standards relied on in this paper (ISO/IEC 27001, 27037, 27042, 27043 and 42001) have been developed through different working groups and on different timetables. Their effective use as a verification suite will require explicit cross-mapping, profile development for the AI-armed-conflict context, and recognition mechanisms across the States that adopt them. National AI risk-management frameworks, including those developed by NIST,[69] are partial analogues but were not designed with verification in mind.

A workable convergence would have three properties. First, an explicit AWS-domain profile of ISO/IEC 42001 that specifies, for AI systems used in armed-conflict contexts, the minimum logging, retention, authentication and validation requirements that a deploying State would have to meet. Second, a documented interface between that profile and the digital-evidence chain-of-custody requirements of ISO/IEC 27037 and 27042, so that the records produced under the AI management system can be handled and interpreted as digital evidence without losing authentication. Third, recognition mechanisms comparable to those used in cybersecurity certification schemes, under which States and international bodies could rely on each other's audit findings when assessing compliance, without requiring full re-audit at every stage.

8.6. Forensic readiness protects honest commanders too

A final point on acceptance. Forensic readiness is sometimes read as a one-sided imposition on militaries: more documentation, more audit, more scope for second-guessing. That reading is incomplete. Records cut both ways. A preserved log may show that the system remained inside a lawful engagement envelope; that the operator received and considered uncertainty warnings; that an override attempt was made; that the harmful classification was the product of an unauthorized post-

---

[68] This observation is the author's; on the adversarial conditions that make machine-generated behaviour difficult to interpret, see NIST, *Adversarial Machine Learning*, above note 43.
[69] NIST AI RMF, above note 21.

deployment modification rather than a baseline design choice; that a commander acted reasonably on the information available at the time. A forensic regime makes the legal inquiry more precise. It identifies what was foreseeable, what was controllable, what was known, what was preserved and what was done. That precision benefits lawful commanders as much as it benefits accountability. A regime that disciplines the inquiry into what actually happened is, in the long run, more protective of professional militaries than a regime in which every adverse outcome is litigated against gauzy declarations.

9. What a Verification Clause Should Require

The argument so far has been general. The minimum content of a verification clause, one that could be included in a future treaty on AI in armed conflict, in a unilateral or multilateral political declaration, in domestic procurement standards, or in military manuals and doctrine, comes to six elements.

First, lifecycle recordkeeping. States should preserve records from training, testing, alignment, procurement, deployment, updates and runtime operations, where those records are necessary to assess compliance with the relevant rule of international law or instrument.

Second, human-control traceability. States should preserve records showing where approval gates were located, who authorized mission parameters, what the operator saw, what warnings were displayed, whether override was practically possible, and whether human intervention occurred or was attempted.

Third, review continuity. A system materially changed through retraining, model update, sensor integration, mission expansion or threshold adjustment should trigger renewed legal and technical review. The review record should identify what changed and whether prior approval remains valid.

Fourth, tamper-evident preservation. Relevant records should be protected through access controls, hash verification, time-stamping, retention rules and chain-of-custody procedures consistent with ISO/IEC 27037 and the Philippine and other domestic electronic-evidence regimes.

Fifth, accountable non-preservation. Failure to preserve required records should carry legal or administrative consequences, especially where the absence of records prevents assessment of civilian harm, human control or command knowledge. Adverse inferences from non-preservation are not novel in evidence law. They operate in domestic criminal and civil proceedings on spoliation, and the same logic has a recognized inter-State form: where one party has exclusive control of the evidence, an international tribunal may resort more liberally to inferences of fact and

circumstantial evidence against it.[70] A deploying State has precisely that exclusive control over the records of its own AI-enabled operations.

A clarification on reach is owed here, because it bears on the mutual mistrust with which this paper opened. Forensic readiness and Evidentiary Custodianship are, in the first instance, things a State does to its own systems, a unilateral self-binding measure whose power to dissolve the mutual dilemma of Section 1 is contingent on reciprocal adoption, which the consensus history of the CCW process shows is the hard part. The adverse-inference logic above bites where a tribunal has jurisdiction, in a domestic prosecution under a statute such as Republic Act No. 9851, or between parties already bound by an instrument. The full mutual case therefore runs through the verification clause proposed below; what unilateral practice does in the meantime is establish the norms, standards and procurement patterns on which any later reciprocal regime would be built.

Sixth, negative disclosure. The procurement specification should require contractors to document, at the design stage, what the system does not log, retain or authenticate, and why. Without this requirement, a State can satisfy every positive logging duty while quietly omitting the categories that would, post-incident, be the most damaging to its position: operator-attention metrics, alignment-stage rollback decisions, override-attempt timestamps, prompt and interface history. After the fact, "we logged everything required" becomes a defense even when the unlogged categories are exactly the ones the legal inquiry needed. Negative disclosure converts that omission from a hidden choice into a procurement-stage commitment. The State has to declare, in advance and on the record, which categories it has decided to render unauditable. That declaration is itself evidence later, and, equally important, it gives reviewers a focal point for pushback before deployment, when leverage still exists. Negative disclosure is the structural counterpart to lifecycle recordkeeping, closing the gap that the positive obligation alone leaves open.

These requirements can and should be scaled. A small loitering munition, a ship-based defensive system, a strategic targeting platform and a foundation-model decision-support tool should not be forced into identical logging patterns. The relevant standard is functional. The greater the autonomy, lethality, operational complexity, civilian-risk profile and distance between human authorization and machine action, the stronger the forensic-readiness duty should be. A low-risk system may require modest records. A system capable of selecting or engaging targets in a civilian-dense environment should require a far richer evidentiary record. The same functional approach should apply to retention. Not every operational log must be preserved indefinitely. Records relevant to legal review, civilian harm, known failure modes, mission authorization, human approval and system action should have defined retention periods. Where civilian harm occurs, preservation should become

[70] International Court of Justice, *Corfu Channel Case (United Kingdom v. Albania)*, Merits, ICJ Reports 1949, p. 4, at p. 18 (a State's exclusive territorial control may permit the victim of a breach a more liberal recourse to inferences of fact and circumstantial evidence).

immediate and mandatory. Where a system is updated after an adverse incident, the prior version should not disappear, otherwise the update itself may erase the evidence needed to understand what went wrong.

The clause architecture also helps reconcile treaty necessity with treaty improbability. If a comprehensive AI-weapons treaty is politically unlikely in the short term, States, regional processes, journals and expert bodies can still develop forensic-readiness norms. Those norms can shape procurement, weapons review, military manuals, political declarations, domestic legislation and future treaty drafting. Huang and Yu's argument that the treaty-making process itself has normative value, even when a treaty is improbable,[71] is consistent with this approach. Forensic readiness gives the process a practical agenda: do not only debate what States should promise, but specify what evidence they must preserve so that the promise can be tested.

10. Conclusion

Forensic readiness is not engineering hygiene. It is the evidentiary architecture without which the legal framework governing AI-enabled force cannot be tested, and therefore cannot be relied on. Existing law (Article 36, Article 28, the precautionary obligations of Additional Protocol I, the Rome Statute, the domestic criminal-law provisions of States like the Philippines) already supplies the substantive constraints. What it does not yet supply is the verification capability that turns those constraints from declarations into evidence. Digital forensics, anchored in ISO/IEC 27001, 27037, 27042, 27043 and 42001, operationalized through procurement, and embedded in international cooperation arrangements, can supply that capability.

Three things follow if the argument is accepted. First, scholarship on AI-enabled force has to take verification seriously as a distinct problem. Discussion of meaningful human control, of appropriate levels of human judgment and of command responsibility cannot proceed as though the elements of these doctrines were self-evidencing. The records do not exist unless they are designed to exist. Second, States that wish to be taken seriously as restrained actors have to design their restraint to be inspectable. A declaration of compliance unbacked by reviewable records is, in the long run, an invitation to infer that the State did not regard the rule it claimed to follow as binding enough to leave a trace. Third, tribunals and inspection bodies have to insist on the records, and where records are absent, to draw the inferences appropriate to that absence. The Prisoner's Dilemma in armed-conflict AI cannot be solved by declarations of restraint: States that wish their restraint to be credible will have to make it inspectable, and States that wish to hold others to account will have to demand that the records exist. Anything else leaves the dilemma's pull intact.

---

[71] Huang and Yu, above note 8.

The proposal does not require new substantive obligations. It requires that existing obligations be designed in advance to be testable. That is what those promises mean now that software has entered the battlespace.