LAWFUL OPERATIONAL SAFEGUARDS IN AI SYSTEMS: SURRENDER RECOGNITION, COMPLIANCE ARCHITECTURE, AND INTERNATIONAL HUMANITARIAN LAW

Giovanni Nardacci*
Founder, Human Flag Association

ABSTRACT

Contemporary discussions concerning AI safeguards frequently frame such mechanisms as discretionary policy choices, corporate alignment preferences, or operational limitations imposed for safety or reputational reasons. This paper proposes a narrower and more legally grounded distinction.

Certain categories of safeguards may instead constitute operational mechanisms relevant to compliance with already-existing obligations under International Humanitarian Law ("IHL"), including obligations reflected in Geneva Convention Common Article 3 and Additional Protocol I, Article 41. In particular, this paper examines *surrender recognition capability* — the technical capacity of a system to identify, preserve, or escalate human surrender and *hors de combat* indicators — as a potentially compliance-relevant safeguard category in military-adjacent and operational AI systems.

The paper does not argue that international law currently mandates any particular AI architecture, nor that all autonomous or semi-autonomous systems are unlawful. It advances a narrower proposition: the systematic exclusion of technically achievable safeguards relevant to surrender recognition and de-escalation may

warrant legal and operational scrutiny under existing humanitarian law frameworks. Under this framework, certain safeguards may warrant analysis not exclusively as optional alignment features, but also as components of *lawful operational design*.

I. INTRODUCTION

Debates concerning artificial intelligence safeguards in operational and military-adjacent environments are often framed in binary terms: more safeguards reduce operational effectiveness, fewer safeguards increase capability. This framing may be incomplete.

International Humanitarian Law imposes obligations concerning the treatment of persons *hors de combat*, including individuals who clearly express surrender or otherwise cease participation in hostilities.[1] These obligations predate modern AI systems by decades and remain legally operative independent of technological evolution.

The emergence of increasingly autonomous decision-support systems, targeting architectures, battlefield intelligence systems, and military-adjacent AI raises a distinct question: whether certain safeguards should be understood not merely as optional policy restrictions, but as operational compliance mechanisms relevant to existing humanitarian obligations.

Existing scholarship has examined extensively whether autonomous systems may comply with IHL obligations as a general matter.[2] The literature has also addressed meaningful human control as a possible compliance criterion.[3] However, it has not systematically addressed whether specific *safeguard categories* may

themselves constitute compliance-relevant operational mechanisms — as distinct from discretionary alignment preferences.[4] This paper addresses that gap.

The paper focuses on *surrender recognition capability* in operational AI systems as an illustrative compliance-relevant safeguard class. It does not address consumer AI, general-purpose language models, or systems without operational or escalation-relevant functions.

The paper advances no claim that current international law explicitly mandates any specific implementation architecture. Instead, it argues that the intentional exclusion of such safeguards from systems capable of operational or battlefield influence may warrant scrutiny under already-existing IHL principles.

This paper proposes an analytical framework for examining the interaction between AI safeguard categories and existing IHL obligations. It does not advance a positive statement of binding customary rules concerning AI architecture, nor does it claim that any single technical implementation is legally mandated. The argument is descriptive-analytical, not prescriptive, and is offered as a contribution to ongoing scholarly and policy discussion (see Methodological Note, *infra*).

Table 1 below illustrates the core distinction proposed in this paper between policy safeguards and lawful operational safeguards, which is developed in full in Section IV.

Table 1: Policy Safeguards vs. Lawful Operational Safeguards

Dimension	Policy Safeguard	Lawful Operational Safeguard
Operational function	Limit capability, reduce reputational or commercial risk	Preserve compliance-relevant information; support lawful engagement obligations

Dimension	Policy Safeguard	Lawful Operational Safeguard
Legal basis	Corporate policy, alignment preference, regulatory positioning	Common Article 3 GC; AP I Articles 41, 57; customary IHL Rule 47
Architecture implication	May be added or removed based on product strategy	Removal may create foreseeable legal and operational risk; warrants documented justification
Verification / audit	Internal policy review; commercial compliance	Independent pre-deploy testing; procurement documentation; feasibility analysis against TRL criteria
Removal consequence	Product or reputational consequence	May engage command responsibility analysis; attribution risk; operational liability

II. EXISTING OBLIGATIONS UNDER INTERNATIONAL HUMANITARIAN LAW

A. Common Article 3

Common Article 3 of the Geneva Conventions establishes minimum protections applicable during armed conflict, including protections for persons who are *hors de combat*. The provision prohibits violence against persons who have laid down their arms or are otherwise no longer participating in hostilities.

The obligation does not depend upon advanced technological conditions. It applies independent of weapon type or operational medium.[1] The ICRC Commentary confirms that the provision was intended to establish a minimum humanitarian floor applicable across all forms of armed conflict regardless of the means employed.[5]

As operational systems become increasingly dependent on automated identification, autonomous classification, machine-assisted targeting, and algorithmic escalation processes, the practical implementation environment changes. The underlying obligation does not.

The legal question that arises is whether system architectures that systematically fail to process or preserve surrender-relevant indicators may create heightened risks of non-compliance with obligations that have been operative for more than seven decades.

B. Additional Protocol I, Article 41

Additional Protocol I, Article 41 elaborates protections for persons recognized as *hors de combat*, including individuals who clearly express an intention to surrender, are incapacitated, refrain from hostile acts, or do not attempt escape.

Article 41 is particularly relevant because it operationalizes recognition. The legal protection is triggered not merely by abstract moral status, but by identifiable conditions and *observable indicators*.[6] The ICRC Commentary explicitly notes that raised arms, abandonment of weapons, and verbal surrender expressions may qualify as indicators regardless of the medium through which they are communicated.[7]

As operational environments increasingly rely on computational mediation, automated sensor interpretation, probabilistic classification, and machine-assisted decision systems, the *capacity to identify such indicators* may acquire direct operational significance.

The question is therefore not whether AI systems become “subjects” of humanitarian law. Rather, the question is whether deployment architectures that

systematically omit surrender-relevant recognition capabilities may undermine the ability of human operators or operational chains to satisfy existing legal obligations.

C. Customary International Humanitarian Law

The prohibition on attacking persons recognized as *hors de combat* is established as customary international law by Rule 47 of the ICRC Customary IHL Study, binding on all parties to a conflict regardless of treaty ratification status.[8] This customary basis confirms that the relevant obligations are not contingent on specific treaty regimes but reflect general principles of the laws of armed conflict applicable across contexts.

D. Command Responsibility

The analysis of lawful operational safeguards intersects with the doctrine of command responsibility under Article 28 of Additional Protocol I and Article 28 of the Rome Statute of the International Criminal Court. These provisions establish that military commanders may incur individual responsibility for crimes committed by forces under their effective control where the commander knew or should have known of the risk, and failed to take necessary and reasonable measures to prevent or repress such acts.

In the context of AI-mediated operational systems, this doctrine may acquire new relevance. Where a commander authorizes deployment of a system that systematically lacks surrender recognition capability — and where such capability was technically achievable and its absence reasonably foreseeable as a compliance risk — questions may arise as to whether the failure to require such safeguards

constitutes a failure to take "necessary and reasonable measures" within the meaning of the command responsibility provisions.

This paper does not argue that current law imposes such liability in any specific case. It observes that as AI systems become more deeply integrated into operational decision chains, the intersection of procurement decisions, system architecture choices, and command responsibility doctrine may warrant increasing attention in both legal scholarship and operational practice.[12]

III. SURRENDER RECOGNITION AS A CAPABILITY CLASS

A. Definition and Technical Feasibility Criteria

For purposes of this paper, *surrender recognition capability* refers to any technical mechanism in an operational AI system designed to: identify surrender expressions; recognize de-escalation indicators; preserve uncertainty during ambiguous encounters; inhibit automatic escalation under uncertainty; escalate ambiguous cases to human review; and maintain operational awareness of *hors de combat* conditions.

The concept does not require perfect recognition. Human operators themselves routinely operate under uncertainty; IHL does not demand infallibility. The relevant inquiry concerns whether foreseeable and technically achievable mitigation mechanisms are systematically excluded.

The term "technically achievable" as used in this paper is not intended as an abstract standard. For purposes of procurement and audit analysis, feasibility may be assessed against verifiable criteria including: Technology Readiness Level (TRL) of relevant sensor and classification components, where TRL 6 or above (system

prototype demonstrated in relevant environment) may indicate sufficient maturity for procurement consideration; acceptable inference latency thresholds for the operational tempo of the system in question; availability of annotated training datasets for surrender-relevant signal categories in representative environments; and capacity for independent validation of recognition performance against defined recall and precision benchmarks. Established technical risk management frameworks such as the NIST AI Risk Management Framework and the IEEE 7000 series on ethical considerations in autonomous systems may provide further reference criteria for structured feasibility assessments.[13]

This framing draws on the principle of precaution in attack reflected in Article 57 of Additional Protocol I, which requires parties to take all *feasible precautions* to avoid or minimize incidental harm.[9] Where surrender recognition may constitute a technically achievable precautionary mechanism against the criteria above, its systematic exclusion may warrant scrutiny under the same analytical framework.

B. Technical Forms

Potential implementation forms may include: multimodal surrender signal recognition; gesture recognition; linguistic surrender detection; uncertainty escalation thresholds; human override escalation pathways; refusal-to-engage constraints under unresolved surrender ambiguity; and probabilistic de-escalation weighting.

This paper intentionally avoids endorsing any single implementation architecture. The legal relevance arises not from a particular model design, but from the *operational function* served.

C. Operational Relevance

In operational environments, escalation errors may occur not only through aggression, but through the *absence* of de-escalation mechanisms. A system optimized exclusively for hostile detection without surrender-sensitive mitigation pathways may increase: false hostile continuation; escalation persistence; failure to preserve surrender ambiguity; and downstream attribution risk.

Under such circumstances, the absence of safeguards does not necessarily increase operational safety or national security. In some environments, the opposite may warrant consideration.[10]

D. Illustrative Operational Scenarios

The following scenarios are illustrative only. They are intended to concretize the abstract compliance risk analysis rather than to assert factual conclusions about any specific system or deployment.

Scenario 1 2014 ISR drone in dense urban environment. An intelligence, surveillance, and reconnaissance (ISR) drone operating in a densely populated urban environment is equipped with an automated threat classification system. The system is optimized to identify hostile actors and flag targets for human-authorized engagement. It does not include any mechanism for detecting or flagging surrender-relevant behaviors — raised hands, discarded weapons, stationary posture following prior movement — as a distinct classification state. In an engagement sequence, a combatant who has ceased hostile activity and raised hands is classified as “still present in threat zone”. The system provides no signal to the human operator that a potentially *hors de combat* condition may exist. The operator, relying on the automated classification, proceeds with engagement authorization.

In this scenario, the absence of a surrender-ambiguity flag in the system architecture may have reduced the human operator's ability to apply Article 41 obligations. Whether a legally relevant failure occurred depends on facts not assumed here. The analytical point is that the system architecture foreclosed a compliance-relevant decision pathway before the human operator was engaged.

Scenario 2 2014 Decision support system for indirect fire. A decision support system for indirect fire operations processes sensor fusion data to provide targeting recommendations to a fire mission authority. The system ranks potential targets by threat probability and recommended engagement priority. It does not include uncertainty escalation logic for scenarios in which available data is insufficient to distinguish active hostile engagement from cessation of hostilities or withdrawal.

A fire mission authority, operating under time pressure and relying on the system's prioritization output, authorizes engagement without independent assessment of *hors de combat* status. Post-engagement review reveals that sensor data available at the time of the recommendation contained indicators consistent with cessation of hostile activity that the system's architecture was not designed to process or flag.

In this scenario, the absence of uncertainty escalation logic for de-escalation-relevant ambiguity may have reduced the practical operability of feasible precaution obligations under Article 57. The procurement decision not to include such logic may warrant analysis under the framework proposed in this paper.

IV. LAWFUL OPERATIONAL SAFEGUARDS

A. Distinguishing Policy Safeguards from Compliance-Relevant Safeguards

Public and regulatory discussions frequently treat all AI safeguards as belonging to a single undifferentiated category. This may obscure an important distinction (see Table 1, above).

Some safeguards are *discretionary policy constraints* reflecting corporate preference, reputational concern, ideological positioning, or product strategy. Others may perform functions directly relevant to legal compliance.

The latter category may be described as *lawful operational safeguards*. Such safeguards are not necessarily mandated in explicit technical form by existing treaties. However, they may constitute foreseeable mitigation mechanisms relevant to operational compliance with preexisting legal obligations.

B. Compliance Architecture

Modern operational systems increasingly depend upon algorithmic classification, probabilistic targeting, autonomous assistance, machine-mediated escalation, and automated operational filtering. As a result, legal compliance increasingly interacts with system architecture.

Under this framework, certain safeguards may function not merely as "ethical add-ons" but as: compliance-preserving mechanisms; operational risk mitigations; uncertainty preservation structures; and escalation-limiting controls.

The potential legal significance of such safeguards arises not from moral aspiration, but from *operational consequence*.

C. Removal as Legal Risk

Where technically achievable safeguards relevant to surrender recognition are intentionally excluded from operational AI systems, foreseeable risks may arise, including: increased escalation under uncertainty; increased probability of failure to recognize *hors de combat* conditions; reduced human review opportunities; and weakened operational restraint pathways.

Under this analysis, removal of such safeguards may warrant legal and operational scrutiny as a potential risk factor rather than a neutral operational choice.[11] This does not imply automatic illegality. It suggests instead that safeguard absence may become relevant to legal and operational evaluation in appropriate contexts — including command responsibility analysis, as discussed in Section II.D.

For audit and procurement purposes, *systematic exclusion* may be evidenced by one or more of the following indicators: procurement specifications that explicitly exclude surrender-relevant signal classes from system requirements; absence of *hors de combat* or de-escalation scenarios in pre-deployment validation protocols; technical documentation or supplier statements indicating that surrender recognition is not a supported or intended capability; or architectural design choices that remove or preclude human escalation pathways in ambiguity-resolving contexts.

V. TECHNICAL LIMITATIONS AND BOUNDARY CONDITIONS

Any credible analysis of surrender recognition as a compliance-relevant capability must acknowledge significant technical limitations. These limitations do not defeat the argument, but they define its proper scope.

A. False Positives, Adversarial Signaling, and Robustness

Recognition systems may generate false positives, incorrectly classifying non-surrender gestures or communications as surrender indicators. In adversarial environments, actors may deliberately exploit surrender recognition mechanisms to create operational pauses, gain tactical advantage, or manipulate escalation dynamics.

The adversarial robustness of gesture and communication recognition systems has been studied extensively in the computer vision and multi-domain operations literature. Research on adversarial examples in deep learning demonstrates that small, deliberately crafted perturbations can cause classification systems to produce incorrect outputs with high confidence.[14] In multi-domain and contested environments, adversarial actors may attempt to generate false surrender signals — through gesture spoofing, acoustic mimicry, or signal injection — to manipulate automated recognition systems. These vulnerabilities are real and must be addressed through structured red-teaming, adversarial scenario testing, and robustness validation as part of any pre-deployment audit process.

However, the existence of adversarial risk does not resolve the underlying legal question. IHL itself operates under conditions of deception and uncertainty and does not suspend obligations because of the risk of false signals. The relevant question is whether the system architecture provides *any* meaningful pathway for surrender-relevant information to influence operational outcomes, and whether the system has been tested against adversarial conditions as part of a structured assurance process.

Recognition systems may also generate *false negatives* — failing to identify genuine surrender indicators due to sensor limitations, model underfitting, or threshold calibration optimized for precision over recall. In IHL terms, a false

negative may directly implicate the precautionary obligation under Article 57 AP I, as it increases the risk of attacking persons *hors de combat*.

The technical trade-off between precision and recall is well-documented in machine learning literature. Where operational systems are calibrated to minimize false positives at the expense of recall, the resulting degradation in surrender-signal detection may warrant specific justification under feasible precaution analysis. This does not imply that perfect recall is required; rather, it suggests that the calibration rationale should be documented and reviewed as part of compliance-relevant assurance processes.

B. Battlefield Ambiguity and Imperfect Recognition Environments

Operational environments present conditions that may substantially degrade recognition performance: low visibility, sensor degradation, partial data, communication interference, and high-tempo decision chains that may limit the time available for uncertainty resolution.

These conditions may constrain the practical implementation of surrender recognition mechanisms in certain operational contexts. They do not, however, eliminate the analytical relevance of the capability class. The principle of feasible precaution under Article 57 of Additional Protocol I is explicitly conditioned on what is *achievable* under the circumstances. Where recognition capability is not feasible given the TRL and operational parameters of the system, the analysis changes accordingly.

The paper therefore does not claim that surrender recognition capability is universally achievable or operationally straightforward. It claims only that its

systematic exclusion from systems where it may be technically feasible — assessed against the criteria in Section III.A — warrants scrutiny.

C. Scope Limitation: Operational Systems Only

The analysis in this paper is confined to AI systems with operational, targeting, escalation, or battlefield-relevant functions. It does not apply to consumer AI, general-purpose language models, logistics systems, or administrative applications.

This scope limitation is analytically significant. The compliance-relevance of surrender recognition capability arises specifically from the operational consequence of its absence in systems capable of influencing kinetic or escalatory outcomes. Systems without such influence fall outside the analytical framework proposed here.

VI. IMPLICATIONS FOR FUTURE GOVERNANCE

A. Autonomous Systems and International Governance

As AI systems become increasingly integrated into operational, military-adjacent, intelligence, and escalation-sensitive environments, governance discussions concerning safeguards will likely intensify. The distinction between discretionary alignment restrictions and lawful operational safeguards may become increasingly relevant in: autonomous systems governance; dual-use AI regulation; military AI procurement; operational accountability analysis; and IHL adaptation debates.

B. European Regulatory Context and Dual-Use Governance

The EU Artificial Intelligence Act (Regulation (EU) 2024/1689) expressly excludes AI systems developed or used exclusively for military or defence purposes from its scope under Article 2(3). The framework proposed in this paper therefore does not depend on, and makes no claim of, direct applicability of the EU AI Act to military AI systems.

However, the European regulatory landscape remains relevant through at least two distinct channels. First, *dual-use AI systems* — systems procured for civilian or governmental purposes that may be deployed in escalation-relevant contexts — may fall within the Act's high-risk classification under Annex III, depending on their specific application. In such cases, the Act's requirements for risk management systems, technical documentation, and human oversight may intersect with the compliance-relevant safeguard analysis proposed here.

Second, European and NATO member state procurement frameworks for defence and security AI systems are subject to national standards, dual-use export controls, and interoperability requirements that may create de facto compliance obligations. The principle of *due diligence* in AI deployment — increasingly reflected in European policy discussions and national procurement guidelines — suggests that responsible deployment may include assessment of whether available safeguard categories relevant to legal compliance have been considered and, where feasible, implemented.

The concept of *foreseeable misuse* as a risk analysis criterion is particularly relevant in this context. Where operational AI systems are deployed in environments where *hors de combat* recognition may be relevant, and where technically achievable safeguards have been systematically excluded, foreseeable misuse analysis may

engage compliance-relevant questions under both IHL and European risk governance frameworks, even where the EU AI Act itself does not directly apply.

C. Procurement Compliance Framework

The procurement framework proposed in this section may be mapped onto existing legal and institutional review processes. For States party to Additional Protocol I, Article 36 reviews of new means and methods of warfare provide a natural procedural anchor for assessing whether surrender recognition capability has been considered as a feasible precaution. Similarly, NATO AI assurance initiatives — including NATO DIANA and the STO AI roadmap — and national frameworks such as the UK MoD's AI approach and US DoD Directive 3000.09 increasingly emphasize structured testing, documentation, and human oversight criteria that align with the checklist elements below. Integrating surrender-relevant safeguard analysis into these established processes may enhance operational accountability without creating parallel or redundant governance structures.

The following framework is offered as a preliminary operational checklist for acquisition authorities, certification bodies, and independent auditors evaluating operational AI systems with escalation-relevant functions. It is not proposed as a legally binding standard, but as a structured analytical reference for feasible precaution analysis under existing IHL principles.

1. Technical Feasibility Assessment

□ Has surrender recognition capability been assessed as technically feasible for this system in its intended operational environment?

□ Has the assessment been documented with reference to TRL criteria, sensor specifications, and latency constraints?

□ Has the assessment been conducted or reviewed independently of the system developer?

2. Exclusion Justification

□ Where surrender recognition capability has been excluded, has the exclusion been documented with specific operational justification?

□ Has the justification addressed whether the exclusion is permanent or context-specific?

□ Has the foreseeable IHL compliance risk created by the exclusion been documented and reviewed?

3. Human Escalation Pathways

□ Does the system architecture include a “pause for verification” pathway triggered by surrender-ambiguous indicators?

□ Is human escalation available before irreversible action in unresolved de-escalation scenarios?

□ Are escalation pathways documented in operational procedures and tested under realistic conditions?

4. Pre-Deployment Audit

□ Has an independent pre-deployment audit been conducted using surrender and de-escalation scenarios?

□ Have adversarial signaling and spoofing scenarios been included in the audit test suite?

□ Has audit documentation been retained for post-deployment operational review?

5. Command Responsibility Documentation

□ Has authorizing command reviewed and acknowledged the IHL compliance risk profile of the system?

□ Is there a documented record of the feasible precaution analysis conducted prior to deployment authorization?

VII. LIMITATIONS OF THE ARGUMENT

This paper does not claim: that current international law prohibits AI systems; that all autonomous systems are unlawful; that surrender recognition is technologically solved; that any single safeguard architecture is legally mandatory; or that operational uncertainty can be eliminated.

The argument is narrower. It proposes only that existing IHL obligations remain operative in AI-mediated operational environments; that certain safeguard categories may serve compliance-relevant operational functions; and that the systematic exclusion of surrender-sensitive safeguards from systems where they may be technically achievable — assessed against verifiable TRL and operational criteria — may warrant legal and operational scrutiny.

This distinction is intended to avoid both technological absolutism and simplistic deregulatory framing. The paper neither advocates categorical AI restrictions nor accepts the proposition that fewer safeguards are necessarily equivalent to greater operational capability or legal safety.

The paper further acknowledges that the translation of IHL principles into technical design specifications raises complex questions that remain unresolved in both legal scholarship and technical practice. The argument presented here is offered as a contribution to that discussion, not as a resolution of it.

KEY TAKEAWAYS FOR PROCUREMENT AND COMMAND STAFF

The following summary is offered for practitioners who may not require the full legal and technical analysis. It does not substitute for the detailed framework set out above.

1. **Function** — Lawful operational safeguards preserve compliance-relevant information. They are not optional ethical add-ons. Their presence or absence may have operational and legal consequences.

2. **Feasibility** — Assess surrender recognition capability against verifiable criteria: TRL ≥6, acceptable latency thresholds, annotated training datasets, independent validation. Document the assessment.

3. **Documentation** — Record exclusions, operational justifications, and feasible precaution analysis. Retain documentation for audit and command review.

4. **Accountability** — Integrate safeguard analysis with Article 28 command responsibility doctrine. Awareness of foreseeable compliance risk is a prerequisite for the ‘reasonable measures’ standard.

5. Verification — Pre-deployment audits should include *hors de combat* scenarios and adversarial robustness testing. Audit documentation should be retained for post-deployment review.

VIII. CONCLUSION

The framing that fewer safeguards necessarily increase operational effectiveness or national security may be incomplete. Certain safeguards may instead function as operational mechanisms whose absence warrants scrutiny under existing humanitarian obligations.

In particular, *surrender recognition capability* in operational AI systems represents a plausible category of lawful operational safeguard. The absence of such capabilities does not necessarily increase operational safety or national security. In some contexts, the systematic exclusion of technically achievable surrender-sensitive architectures — where technically achievable is assessed against the TRL and operational criteria proposed in Section III.A — may warrant consideration of foreseeable escalation, non-compliance, and operational liability risks, including potential implications for command responsibility doctrine.

The purpose of this paper is not to prohibit AI systems, nor to mandate any single implementation architecture. Its narrower objective is to introduce a legal distinction — between discretionary alignment preferences and compliance-relevant safeguard categories — that may prove analytically useful in future governance, procurement, and legal accountability discussions concerning AI systems in operational environments.

If the analysis presented here is correct, the relevant question in those discussions is not whether AI systems should have fewer or more safeguards as a

general matter. The relevant question is *which safeguards* may perform compliance-relevant functions under existing law, what verifiable criteria determine their technical feasibility, and how the sequence *technical feasibility → documented justification → command awareness → operational accountability* may structure future governance, procurement, and legal review.

METHODOLOGICAL NOTE

This paper proposes an analytical framework rather than a positive statement of existing customary obligations concerning AI architecture. The argument is offered as a contribution to ongoing scholarly and policy discussion. It does not purport to resolve disputed questions of treaty interpretation, nor to establish binding standards of conduct.

The illustrative scenarios presented in Section III.D are analytical constructs. They are intended to demonstrate how the proposed framework might apply to specific operational configurations, not to describe or evaluate any existing system. References to procurement frameworks and audit criteria in Section VI.C are offered as analytical reference points and do not constitute legal advice.

The author welcomes engagement, critique, and further development of the framework by scholars, practitioners, and policymakers working in international humanitarian law, autonomous systems governance, and defence procurement.

Cite as: Giovanni Nardacci, ‘Lawful Operational Safeguards in AI Systems: Surrender Recognition, Compliance Architecture, and International Humanitarian

Law' (2026) (proposing distinction between discretionary safeguards and compliance-relevant operational safeguards under existing IHL principles).

* Giovanni Nardacci is the founder of Human Flag Association, a Switzerland-based humanitarian technical initiative focused on civilian protection standards and IHL compliance architectures in AI-enabled operational environments. UNGM Partner No. 8128. humanflag.org. The author prepared this paper independently; no funding was received from any party with a commercial or governmental interest in the subject matter.

NOTES

1. Jean S. Pictet (ed.), *Commentary on the Geneva Conventions of 12 August 1949, Vol. IV* (ICRC, 1958) 39–40. See also Jean-Marie Henckaerts & Louise Doswald-Beck, *Customary International Humanitarian Law, Vol. I: Rules* (ICRC/Cambridge University Press, 2005) Rule 47, 164–169.

2. See, e.g., Michael N. Schmitt & Jeffrey S. Thurnher, 'Out of the Loop: Autonomous Weapon Systems and the Law of Armed Conflict' (2013) 4 *Harvard National Security Journal* 231; Kenneth Payne, *I, Warbot: The Dawn of Artificially Intelligent Conflict* (Hurst, 2021); Human Rights Watch & IHRC, *Losing Humanity: The Case Against Killer Robots* (2012). For more recent analysis, see ICRC, *IHL and New Technologies: Overview* (ICRC, 2021); SIPRI, *Artificial Intelligence in Military Systems and its Implications for the Arms Control Regime* (SIPRI Policy Paper No. 61, 2021).

3. UNIDIR, *The Weaponization of Increasingly Autonomous Technologies: Considering How Meaningful Human Control Might Move the Discussion Forward* (UNIDIR, 2014); Article 36, *Killer Robots: UK Government Policy on Fully Autonomous Weapons* (2013). On human-machine teaming in targeting, see US Department of Defense Directive 3000.09 (Autonomy in Weapon Systems, updated 2023).

4. The distinction proposed here differs from the debate on meaningful human control. The question is not whether a human is in the loop, but whether the system architecture itself contains mechanisms capable of preserving compliance-relevant information — including surrender and de-escalation signals — for human or automated review. This is a distinct analytical question that the existing literature has not systematically addressed.

5. International Committee of the Red Cross, *Autonomous Weapon Systems and International Humanitarian Law: A Guide to the Issues* (ICRC, 2014); ICRC, *IHL and New Technologies: Overview* (ICRC, 2021); ICRC, *Autonomous Weapons Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons* (ICRC Expert Meeting Report, 2023).

6. Yves Sandoz, Christophe Swinarski & Bruno Zimmermann (eds.), *Commentary on the Additional Protocols of 8 June 1977 to the Geneva Conventions of 12 August 1949* (ICRC/Martinus Nijhoff, 1987) paras. 1593–1621.

7. Sandoz et al., *supra* note 6, para. 1602.

8. Henckaerts & Doswald-Beck, *supra* note 1, Rule 47, 166.

9. Additional Protocol I, Art. 57. See also ICRC, *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law* (ICRC, 2009) 77–82.

10. This observation is consistent with scholarship on escalation dynamics in autonomous systems. See Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (W.W. Norton, 2018) ch. 14. See also RAND Corporation, *Human-Machine Teaming for Future Ground Forces* (RAND, 2020).

11. The notion of foreseeable risk as a basis for operational legal analysis is well established in IHL. See ICRC Interpretive Guidance, *supra* note 9, 77–82.

12. On command responsibility and autonomous systems, see Schmitt & Thurnher, *supra* note 2, 260–265; Marco Sassoli, 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified' (2014) 90 *International Law Studies* 308. See also Human Rights Watch & IHRC, *Mind the Gap: The Lack of Accountability for Killer Robots* (HRW, 2015).

13. National Institute of Standards and Technology, *AI Risk Management Framework (AI RMF 1.0)* (NIST AI 100-1, 2023). IEEE, *IEEE 7000-2021: Model Process for Addressing Ethical Concerns During System Design* (IEEE, 2021). On Technology Readiness Levels in defence procurement, see NATO, *Technology Readiness Levels in NATO Science and Technology* (NATO STO, 2019). On AI assurance in defence systems, see UK Ministry of Defence, *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence* (MoD, 2022).

14. On adversarial examples in deep learning, see Christian Szegedy et al, 'Intriguing Properties of Neural Networks' (2014) *ICLR 2014*; Ian J Goodfellow, Jonathon Shlens and Christian Szegedy, 'Explaining and Harnessing Adversarial Examples' (2015) *ICLR 2015*; Nicholas Carlini and David Wagner, 'Towards Evaluating the Robustness of Neural Networks' (2017) *IEEE Symposium on Security and Privacy* 39

SELECT BIBLIOGRAPHY

Primary Sources — International Legal Instruments

Geneva Conventions of 12 August 1949, Common Article 3.

Protocol Additional to the Geneva Conventions of 12 August 1949 (Protocol I), 8 June 1977, Articles 41, 57.

Rome Statute of the International Criminal Court, 17 July 1998, Article 28.

Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 (EU Artificial Intelligence Act), Articles 2(3), Annex III.

ICRC Materials

Henckaerts, Jean-Marie & Doswald-Beck, Louise. *Customary International Humanitarian Law, Vol. I: Rules.* ICRC/Cambridge University Press, 2005.

ICRC. *Autonomous Weapon Systems and International Humanitarian Law: A Guide to the Issues.* ICRC, 2014.

ICRC. *Autonomous Weapons Systems: Implications of Increasing Autonomy in the Critical Functions of Weapons.* ICRC Expert Meeting Report, 2023.

ICRC. *Interpretive Guidance on the Notion of Direct Participation in Hostilities under International Humanitarian Law.* ICRC, 2009.

Pictet, Jean S. (ed.). *Commentary on the Geneva Conventions of 12 August 1949, Vol. IV.* ICRC, 1958.

Sandoz, Yves, Swinarski, Christophe & Zimmermann, Bruno (eds.). *Commentary on the Additional Protocols of 8 June 1977.* ICRC/Martinus Nijhoff, 1987.

Academic Literature

Carlini, Nicholas & Wagner, David. 'Towards Evaluating the Robustness of Neural Networks.' (2017) *IEEE Symposium on Security and Privacy* 39.

Goodfellow, Ian J., Shlens, Jonathon & Szegedy, Christian. 'Explaining and Harnessing Adversarial Examples.' (2015) *ICLR 2015*.

Payne, Kenneth. *I, Warbot: The Dawn of Artificially Intelligent Conflict.* Hurst, 2021.

Sassoli, Marco. 'Autonomous Weapons and International Humanitarian Law: Advantages, Open Technical Questions and Legal Issues to be Clarified.' (2014) 90 *International Law Studies* 308.

Scharre, Paul. *Army of None: Autonomous Weapons and the Future of War*. W.W. Norton, 2018.

Schmitt, Michael N. & Thurnher, Jeffrey S. ‘Out of the Loop: Autonomous Weapon Systems and the Law of Armed Conflict.’ (2013) 4 *Harvard National Security Journal* 231.

Szegedy, Christian et al. ‘Intriguing Properties of Neural Networks.’ (2014) *ICLR 2014*.

Policy, Institutional and Technical Documents

Article 36. *Killer Robots: UK Government Policy on Fully Autonomous Weapons.* 2013.

Human Rights Watch & IHRC. *Losing Humanity: The Case Against Killer Robots.* 2012.

Human Rights Watch & IHRC. *Mind the Gap: The Lack of Accountability for Killer Robots.* 2015.

IEEE. *IEEE 7000-2021: Model Process for Addressing Ethical Concerns During System Design.* IEEE, 2021.

NATO STO. *Technology Readiness Levels in NATO Science and Technology.* NATO, 2019.

NIST. *AI Risk Management Framework (AI RMF 1.0).* NIST AI 100-1, 2023.

RAND Corporation. *Human-Machine Teaming for Future Ground Forces.* RAND, 2020.

SIPRI. *Artificial Intelligence in Military Systems and its Implications for the Arms Control Regime.* SIPRI Policy Paper, 2021.

UK Ministry of Defence. *Ambitious, Safe, Responsible: Our Approach to the Delivery of AI-Enabled Capability in Defence.* MoD, 2022.

US Department of Defense. *Directive 3000.09: Autonomy in Weapon Systems* (updated 2023).

UNIDIR. *The Weaponization of Increasingly Autonomous Technologies.* UNIDIR, 2014.