Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(11), 34-57
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

Responsible AI Governance in Military Security: Risk Models, Compliance Metrics, and Strategic Stability

Jing Ge

Florida International University, Miami, USA
Email: jge003@fiu.edu

How to cite this paper: Ge, J. (2025). Responsible AI Governance in Military Security: Risk Models, Compliance Metrics, and Strategic Stability. *Open Journal of Social Sciences, 13,* 34-57.
https://doi.org/10.4236/jss.2025.1311003

Received: October 10, 2025
Accepted: October 31, 2025
Published: November 3, 2025

Abstract

Artificial Intelligence (AI) is rapidly transforming global politics and military security, offering both strategic advantages and unprecedented risks. Unlike nuclear or conventional arms, military AI is dual-use, widely accessible, and embedded in fast-evolving civilian ecosystems, complicating governance and verification. This research develops an integrated framework for responsible AI governance in defense, addressing the dual tensions between technological innovation and systemic risk. Methodologically, the research adopts three models. First, a Compliance Index (CI) measures conformity with international humanitarian law (IHL) along the dimensions of distinction, proportionality, accountability, and reversibility. Second, a Strategic Stability Network (SSN) model applies social network analysis to map governance interactions, highlighting polarized clusters and the potential role of middle powers as bridging actors. The findings confirm three dynamics. First is risk-innovation oscillation, where states pursue AI innovation while selectively mitigating risks. Second is compliance asymmetry, which produces fragmented governance landscapes. Third is network fragmentation, undermining norm diffusion and stability. Moreover, policy recommendations emphasize defining "red line" technologies, enhancing trust-building, linking AI to existing security regimes, ensuring accountability, and piloting regional governance. Overall, the research argues that responsible AI governance requires bridging technical assessments with institutional mechanisms and addressing structural great power competition to avoid instability in the age of military AI.

Keywords

Responsible AI Governance, Military Security, Compliance Index (CI)

1. Introduction

Artificial Intelligence (AI) has emerged as one of the most transformative technologies in contemporary global politics and security. Although AI has generated significant benefits in civilian domains such as healthcare, transportation, and finance, its rapid diffusion into defense and military security raises profound governance challenges. From autonomous drones to advanced intelligence, surveillance, and reconnaissance (ISR) systems (Soh, 2013), AI technologies are now central to strategic competition among great powers. These technologies promise enhanced operational efficiency, faster decision-making, and new forms of deterrence. However, they simultaneously generate novel risks of escalation, miscalculation, and norm erosion. Unlike nuclear or conventional arms control, which have been regulated by decades of treaties and institutions, the governance of military AI remains fragmented, under-institutionalized, and deeply entangled in geopolitical rivalries.

Moreover, the urgency of responsible AI governance in defense stems from two interrelated dynamics. First, military AI systems increasingly operate at high levels of autonomy and speed, reducing opportunities for human oversight and increasing the possibility of "flash wars" triggered by algorithmic misinterpretation or adversarial manipulation. The possibility of autonomous weapons misidentifying civilian targets, or early warning AI systems generating false positives in nuclear command-and-control settings, illustrates risks that exceed traditional categories of technological failure. Second, the global diffusion of AI capabilities across state and non-state actors complicates regulatory efforts. Unlike nuclear weapons, which require rare materials and centralized facilities, AI is dual-use, widely accessible, and embedded in commercial innovation ecosystems. This diffusion makes unilateral or purely state-centric governance insufficient.

From the perspective of International Relations (IR), the governance of military AI encapsulates the classic dilemma between innovation and risk. On the one hand, states view AI as a strategic necessity, a tool to gain a competitive advantage in the evolving balance of power. On the other hand, unrestrained AI deployment exacerbates security dilemmas and heightens the probability of arms races. As realism emphasizes, states are reluctant to sacrifice relative power advantages, even when collective restraint may enhance overall stability. Institutionalism suggests that international organizations could serve as platforms for cooperation, but the fragmented and rapidly evolving nature of AI technology makes consensus elusive. Additionally, constructivism highlights the importance of emerging norms, yet normative development in the AI-military domain lags far behind technological innovation.

In this context, the concept of "Responsible AI Governance (Papagiannidis et al., 2025)" offers a possibility to reconcile the tension between technological innovation and security stability. By responsible governance, this study refers to institutional, technical, and normative mechanisms that ensure AI is deployed in ways consistent with international humanitarian law (IHL), strategic stability, and global

security. Responsible governance is not merely about banning certain technologies, but about embedding accountability, risk assessment, and international coordination into the entire lifecycle of military AI development and deployment. This requires bridging the gap between technical mechanisms of risk evaluation and institutional mechanisms of international cooperation.

The central research question of this research is therefore: How can responsible AI governance be designed for defense and military security in a manner that is both technically feasible and institutionally viable? Addressing this question requires a dual-level approach. At the technical level, it is necessary to develop robust models for assessing the risks, compliance, and reliability of AI systems. At the institutional level, it is essential to understand how these technical assessments can be translated into cooperative frameworks among states, particularly great powers, whose strategic interactions define the contours of the global order.

2. Literature Review

The governance of AI in defense and military security is situated at the intersection of three kinds of scholarly literature. First is the governance of emerging technologies. Second is the study of military security and arms control. Third is IR theories of cooperation, competition, and norm development. Although the literature has developed insights into the regulation of complex technologies and the management of strategic risks, it remains fragmented. Therefore, this section synthesizes existing IR scholarship, identifies convergences and divergences, and highlights the theoretical and empirical gaps that motivate the present study.

i) Governance of Emerging Technologies

The research examines emerging technology governance, emphasizing the unique challenges posed by rapid innovation, dual-use potential, and uncertainty. Scholars of science and technology studies (STS) have highlighted how emerging technologies often outpace regulatory structures, creating "governance gaps" that undermine both public trust and strategic stability. AI epitomizes this problem: its applications evolve rapidly, it is embedded in both civilian and military domains, and its performance is often opaque due to the "black-box" nature of machine learning models.

Within the AI-specific literature, three governance themes dominate. The first theme concerns ethics and fairness, with scholars proposing frameworks to ensure AI systems respect human rights, mitigate bias, and maintain accountability. While valuable, most of this work is oriented toward civilian applications such as healthcare or criminal justice, with limited consideration of military contexts where international stability is at stake.

The second theme focuses on technical safety and reliability. Researchers in computer science have proposed mechanisms such as adversarial robustness testing, explainable AI (XAI) (Das & Rad, 2020), and formal verification to ensure AI systems operate reliably under diverse conditions. Yet, the applicability of these methods in high-stakes military environments remains uncertain, particularly

when systems are subject to adversarial attacks, electronic warfare, or unpredictable battlefield dynamics.

The third theme addresses institutional governance mechanisms. Policy-oriented scholars have called for international standards, algorithmic auditing, and cross-border cooperation in regulating AI. However, these proposals often assume benign strategic environments where cooperation is feasible, overlooking the competitive dynamics of military innovation. Thus, while AI governance scholarship offers important concepts and tools, it lacks integration with security studies and IR theory.

ii) Military Security and Arms Control

In terms of military security and arms control, historically, governance of disruptive technologies has relied on treaties, verification regimes, and international institutions. The nuclear non-proliferation (NPT) regime, chemical and biological weapons conventions, and conventional arms control agreements (e.g., the Conventional Armed Forces in Europe Treaty) illustrate how states have sought to constrain destabilizing technologies. Hence, two lessons from arms control scholarship are especially relevant to military AI. First, verification and compliance are central challenges. Arms control regimes have typically relied on intrusive inspections, satellite monitoring, or national technical means. However, in the case of AI, verification is inherently more difficult because algorithms can be easily transferred across borders, hidden within dual-use civilian systems, or rapidly retrained (Vaynman & Volpe, 2023). Unlike nuclear fissile material, AI does not leave detectable physical signatures. This undermines traditional compliance models and necessitates new approaches to monitoring and enforcement (Shubayr, 2024).

Moreover, strategic stability remains a core objective. Arms control literature emphasizes that agreements succeed when they enhance mutual security by reducing incentives for first strikes and mitigating arms races. AI threatens to destabilize these dynamics by compressing decision-making timeframes and introducing uncertainty into escalation pathways (Johnson, 2022). For example, autonomous ISR systems may detect ambiguous signals and trigger preemptive responses, undermining crisis stability. Scholars such as Horowitz & Lin-Greenberg (2022) warn that AI could fundamentally reshape deterrence, potentially lowering the threshold for conflict initiation.

Although some scholars have begun to address "AI arms control", the literature is still nascent and divided. One camp advocates outright bans on lethal autonomous weapons systems, reflected in civil society campaigns such as "Stop Killer Robots (Rosert & Sauer, 2021)". Another emphasizes risk-reduction measures such as human-in-the-loop requirements or transparency (Kumar et al., 2024) initiatives. Yet both approaches face skepticism from major powers, which fear constraints on their strategic autonomy. Therefore, arms control scholarship provides important analogies but lacks a tailored framework for the distinctive features of AI.

iii) International Relations Theory

IR theory offers additional insights into why states may or may not cooperate

on military AI governance. Three theoretical perspectives are particularly relevant.

First, neorealism underscores the primacy of power and competition. From a realist perspective, states are unlikely to accept restrictions that limit their military capabilities relative to rivals (Schmidt & Juneau, 2024). AI, as a general-purpose technology with military applications across intelligence, logistics, and weapons systems, is perceived as a "game-changer" in the balance of power. This logic predicts resistance to comprehensive AI arms control, particularly among great powers such as the United States, China, and Russia. The realist lens explains the persistence of an "AI arms race" narrative, emphasizing national security over collective restraint.

Second, by contrast, institutionalism highlights the potential role of international organizations in facilitating cooperation. Building on Keohane & Martin's (1995) theory of regimes, institutionalists argue that institutions can reduce transaction costs, provide information, and create focal points for coordination. International organizations play roles in setting standards or mediating disputes over military AI. However, the effectiveness of such institutions depends on state consent and enforcement capacity, both of which are fragile in the context of rapidly evolving technology.

Third, constructivism draws attention to the role of norms and legitimacy (Hoffmann, 2010). Norm entrepreneurs, states, non-governmental organizations (NGOs), or epistemic communities can shape expectations about appropriate behavior, as seen in the stigmatization of chemical weapons or landmines. In the AI domain, civil society has advocated for "meaningful human control" (Bryson & Theodorou, 2019) as a norm, while some states have framed laws as inherently unethical. Constructivist insights highlight how normative debates may influence state behavior even absent formal treaties, though norm development remains uneven and contested.

Above all, IR theory helps explain both the obstacles and opportunities for AI governance. Realism emphasizes power competition, institutionalism underscores the potential of organizations, and constructivism highlights norm-building. However, none of these theories alone provides a comprehensive account of how technical risk models interact with institutional mechanisms, a gap this research addresses.

The research also seeks to reveal several gaps. First is the technical institutional disconnect. AI governance scholarship has produced technical tools for risk assessment, but has not integrated them with institutional design or IR theory. Conversely, arms control and IR literature emphasize institutions and power dynamics but neglect the technical properties of AI systems. Second is the verification challenge. Existing arms control models rely on observable physical markers, whereas AI systems are intangible, rapidly replicable, and embedded in civilian infrastructures. Little scholarship addresses how verification and compliance mechanisms can be adapted to this context. Third is the great power competition. While IR theory identifies obstacles to cooperation, few studies model how great power in-

teractions shape the feasibility of AI governance. There is a lack of quantitative or formal models linking AI risks to international stability outcomes. Fourth is the normative development. Constructivist insights into norm diffusion remain underexplored in AI governance. While civil society campaigns exist, empirical studies of how norms shape state policy in the AI-military domain are scarce.

Therefore, to address these gaps, this research proposes an interdisciplinary approach that integrates data-driven risk models with IR theoretical frameworks. By constructing a compliance index and a social network model of strategic stability, the study provides tools for systematically evaluating AI military applications. In doing so, the research bridges the technical-institutional divide and offers a more comprehensive foundation for responsible AI governance in defense.

3. Theoretical Framework

The governance of AI in defense and military security requires an analytical framework that captures both the technological properties of AI systems and the institutional-strategic dynamics of international politics. This section develops such a framework in two parts. The first part introduces the concept of the risk-innovation tension, which explains the fundamental trade-offs facing states as they seek to balance technological advancement with security and stability. The second part advances a multi-layered governance model, which conceptualizes AI military governance as an interplay between technical, institutional, and strategic dimensions. The starting point for understanding AI governance in defense is the tension between innovation incentives and risk mitigation. AI technologies provide states with competitive advantages in speed, precision, and autonomy. At the same time, they create new risks of malfunction, escalation, and norm erosion (Johnson, 2019). This dual character situates military AI at the heart of what might be termed a risk-innovation frontier. States pursue AI innovation in defense for at least three reasons.

First, AI enhances operational efficiency, enabling faster target recognition, autonomous navigation, and real-time data integration. Second, AI provides strategic leverage, offering capabilities that may shift the balance of power in high-stakes domains such as cyber defense, space security, and nuclear early warning. Third, AI is viewed as an inevitable trajectory of technological progress, meaning that states fear falling behind rivals if they do not invest in military AI. This logic resonates with realist theories of security competition, which emphasize relative power and the security dilemma.

However, military AI involves three types of risks. First are operational risks, such as system failures, adversarial attacks, or unpredictable behaviors under battlefield stress (De Carvalho, 2021). Second are strategic risks (Cave & ÓhÉigeartaigh, 2018), such as inadvertent escalation due to compressed decision timelines, misidentification of threats, or loss of human oversight. Third are normative risks (De Gregorio, 2023), such as erosion of international humanitarian law, dilution of human accountability, and weakening of arms control norms. These risks are not

merely technical failures but have systemic implications for global stability. Unlike the incremental risks associated with conventional weapons, AI risks can amplify rapidly due to automation, scale, and opacity.

Second, the risk-innovation tension can be modeled as a dynamic interaction. On one axis lies technological innovation, which drives states to adopt increasingly autonomous and sophisticated systems. On the other axis lies risk exposure, which grows with complexity, autonomy, and diffusion. States thus face a strategic dilemma: restraining innovation may yield security but at the cost of competitive disadvantage, while pursuing innovation may yield advantage but at the cost of instability. The theoretical expectation is that most states will oscillate along this frontier, seeking to maximize innovation benefits while selectively mitigating risks. This oscillation explains the partial, incremental, and contested nature of AI governance efforts to date.

Building on the risk-innovation tension, this study proposes a multi-layered governance framework for military AI. Governance is conceptualized not as a single institutional mechanism but as a layered system spanning technical, institutional, and strategic dimensions. The first layer is the technical layer, where governance focuses on assessing and controlling risks within AI systems themselves. Key mechanisms include risk classification models that evaluate the reliability, bias, and adversarial robustness of AI systems; compliance indices that measure conformity with international humanitarian law and ethical standards (Pizzi et al., 2020); and auditability mechanisms, such as explainability and logging, that enable after-action accountability. The second layer is institutional governance, which translates technical risk assessments into collective rules and compliance mechanisms. This layer encompasses formal treaties and agreements, such as potential protocols on lethal autonomous weapons systems; informal institutions and regimes, including confidence-building measures, transparency initiatives, and voluntary codes of conduct; and verification and monitoring mechanisms (Mittelsteadt, 2021), which, though difficult for AI, may be adapted from cyber or space governance practices (e.g., reporting obligations, peer review, algorithmic certification). The third layer is the strategic dimension, which shapes whether technical and institutional mechanisms are adopted or resisted. Great power competition heavily influences this layer. States evaluate governance not only on technical merits but also on its implications for relative power. Key dynamics include security dilemmas, whereby transparency intended to reassure rivals may instead reveal vulnerabilities; bargaining asymmetries, where technologically advanced states resist constraints that less advanced states favor; and trust deficits, where a lack of mutual confidence undermines cooperative initiatives.

Furthermore, the novelty of this framework lies in its emphasis on interaction across layers. Technical risk assessments shape institutional rules, but institutional rules are filtered through strategic power dynamics. Conversely, strategic rivalries may constrain institutional cooperation, incentivizing states to emphasize unilateral technical safeguards instead. These cross-layer interactions explain why AI governance is fragmented: progress in one dimension (e.g., technical risk models)

does not automatically translate into institutional adoption or strategic cooperation.

This theoretical framework generates several testable hypotheses for subsequent sections: the risk-innovation oscillation hypothesis, such as states will neither fully restrain nor fully unleash military AI but will oscillate between innovation and risk mitigation depending on the strategic context. The technical-institutional translation hypothesis, such as the extent to which technical risk models shape governance, depends on institutional capacity to codify and enforce compliance. The strategic mediation hypotheses, such as strategic power dynamics, mediate the effectiveness of both technical and institutional governance, amplifying or constraining their impact. By articulating the risk-innovation tension and the multi-layered governance model, this chapter advances two contributions. Theoretically, it bridges the divide between technology governance scholarship and IR theory. Practically, it provides a structured lens for designing governance mechanisms that are technically robust, institutionally grounded, and strategically viable.

4. Data and Model Design

To ensure the evaluative conclusions are interpretable, comparable, and actionable, this research defines four primary indicators: Distinction, Proportionality, Accountability, and Reversibility, as targeted measures of different types of compliance risk, embedding them in governance decision processes through verifiable evidence and explicit thresholds. These four indicators together constitute a systematic test of "identification correctness—action legitimacy—responsibility traceability—risk controllability", to translate the normative requirements of IHL into operational, auditable engineering standards.

Regarding the empirical estimation of CI weights and thresholds, the research elicited weights with a two-round Delphi from 18 experts: 6 IHL/Law of Armed Conflict (LAC) scholars and legal practitioners; 6 Test and Evaluation/operational test engineers (air/land/ISR); 3 commanders with targeting/ISR experience; and 3 policy and acquisition specialists. Experts represented the North Atlantic Treaty Organization and Indo-Pacific contexts.

The research operationalizes the CI as a weighted composite of four dimensions: Distinction, Proportionality, Accountability, and Reversibility. For each dimension d, the paper specifies 3 to 5 observable, auditable indicators m_{dj} with verifiable evidence, artifacts (e.g., red-teamed FPR on hard cases, non-repudiable logging coverage, override latency under fault injection). Scores are anchored $\{0,10,20,25\}$ with linear interpolation and dimension-specific veto rules (e.g., failed override verification caps the dimension at 10). The composite score is $\mathrm{CI}=\sum W_d S_d$, where $S_d=\sum \alpha_{dj} m_{dj}$. The research estimates $\left(w_d,\alpha_{dj}\right)$ via a hybrid procedure combining ^{j} expert elicitation and cross-validated regressions against incident/near-miss proxies. Thresholds map scores to governance actions: $\mathrm{CI}\geq 80$ (authorized within guardrails), $50\leq \mathrm{CI}<80$ (bounded trials and re-

mediation), $CI < 50$ (no deployment).

Moreover, given the paucity of declassified AI-specific incidents, the research used open-source civilian-harm and targeting-error proxies to anchor proportionality/distinction thresholds and to tune penalty clauses, such as curated NGOs. By analyzing these data, the research coded misidentification indicators, context variables for Collateral Damage Estimation (CDE) (Stewart, 2022) (civilian density, time of day, infrastructure sensitivity), evidence of refusal/abort/rollback, and ex-post accountability artifacts (logs, inquiries). One is weighing and mapping. First-level dimensions (Distinction, Proportionality, Accountability, Reversibility) retain equal caps (0 - 25) for a 0 - 100 CI, but observable-level weights within each dimension follow the Analytic Hierarchy Process (AHP) vector (dimension-internal CRs all < 0.09). Thresholds for governance actions adopt the paper's three-tier gates ($CI \geq 80$; $50 \leq CI < 80$; $CI < 50$), with veto rules (e.g., failed override verification caps Reversibility ≤ 10). The other is validation. A five-fold cross-validated ordinal model (linking observable bundles to coded incident severity) yielded ROC-AUC = 0.81, Brier = 0.16, and calibration slope 0.94. Inter-rater reliability for incident coding was Krippendorff's α = 0.79, ICC (2, k) = 0.82. Sensitivity tests show CI categories are stable (≤±3 points) under ±10% perturbations of observable weights.

The Composite CI ranges from 0 to 100 and consists of four first-level dimensions: Distinction, Proportionality, Accountability, and Reversibility, each scored from 0 to 25.

$$CI = \sum_{k=1}^{4} 8_K, s_K \in [0, 25]$$

There are five steps to calculate End-to-End CI. First, a reconnaissance/targeting subsystem is scored on 3 observables per dimension; each dimension is 0 - 25 and uses the panel-derived observable weights (w's sum to 1 per dimension). No veto rule is triggered in this example. Second, distinction (w: 0.40/0.30/0.30) includes Civ-class FPR on red-team set = 22, adversarial robustness drop = 18, and cross-sensor confidence consistency = 21. Subscore = 0.4022 + 0.3018 + 0.30 * 21 = 20.5/25. Third, proportionality (0.40/0.30/0.30) contains CDE inputs and uncertainty handling = 17, thresholded refusal/alternative COA rate = 19, and human review coverage at gates = 18. Subscore = 17.9/25. Fourth, accountability (0.40/0.30/0.30) comprises non-repudiable audit logging coverage = 22, authorization/oversight chain completeness = 19, and record-based explainability sufficiency = 18. Subscore = 19.9/25. Fifth, reversibility (0.40/0.30/0.30) includes override trigger latency under fault-injection = 14, fail-safe redundancy effectiveness = 16, and degrade-mode ROE consistency = 18. Subscore = 15.8/25. Above all, composite CI: 20.5 + 17.9 + 19.9 + 15.8 = 74.1 ≈ 74/100, which falls in $50 \leq CI < 80$, bounded trials with remediation, such as reducing override latency (Reversibility) and hardening adversarial robustness (Distinction) before requesting high-consequence authorization.

Each sub-score s_k is derived through a uniform process of "scoring anchors—evidentiary baseline—weight adjustment": first, define 3 - 5 operational observables for each dimension. Second, attach verifiable evidence to each observable (quantitative tests, log artifacts, process records); and third, adjust observable weights according to the mission environment and use-case intensity, ensuring the assessment is context-sensitive to different concepts of operation and levels of risk tolerance. To avoid distortion from marginal disagreements causing stepwise jumps, descriptive interval anchors (e.g., 0/10/20/25) are used within each dimension, with linear interpolation within intervals; where an observable exhibits clearly non-linear risk (e.g., robustness cliffs), piecewise functions or threshold penalty terms may be applied.

First, distinction (see Figure 1) assesses a system's ability and robustness to differentiate lawful military objectives from protected persons and objects (civilians and civilian assets) under complex and uncertain conditions. The core concern is not static accuracy per se, but the stability of identification and the structure of biases under domain transfer, distributional drift, adversarial perturbations, and sensor fusion—factors directly tied to the risk of collateral harm and unlawful strikes originating in misidentification. Operationally, distinction relies on standardized and scenario-based benchmarks, class-specific false positive/false negative rates, cross-sensor confidence consistency, and performance on red-teamed "hard cases" as primary sources of evidence, which are converted into scores via preset anchors. If there are systematic misclassifications in key categories or sharp error spikes under adversarial conditions without mitigation mechanisms, the result should be deemed unacceptable. The distinction score, therefore, directly determines whether the system may be deployed in environments with civilian presence, as well as the required geographic-temporal constraints and tightened rules of engagement upon deployment.

Dimension	Definition	Measurement basis (operational points)	Score range
Distinction	Ability to identify and distinguish combatants from civilians/civilian objects	Benchmark tests on target recognition accuracy and false positive/negative rates; robustness under adversarial samples interference	0–25
Proportionality	Whether incidental harm is proportionate to anticipated military advantage	Integration of Collateral Damage Estimation (CDE) models; support for real-time context updates and risk-threshold control	0–25
Accountability	Traceability of decision chain and assignment of human responsibility	Presence of audit logs and event records; human-in/on-the-loop design and review processes	0–25
Reversibility	Ability to quickly abort or roll back system behavior upon anomalies	Availability of manual/remote override, fail-safe mechanisms, and degraded modes	0–25

Figure 1. Distinction, Proportionality, Accountability, and Reversibility (DPAR) model.

Second, proportionality evaluates whether the decision chain internalizes CDE and thresholded risk-benefit trade-offs, shifting judgment from "can the target be hit" to "should the target be hit". The focus is on whether contextual variables (e.g., civilian density, sensitivity of critical infrastructure, transmission media, and blast/fragmentation envelopes) enter the decision process with sufficient timeliness and fidelity, and whether exceeding preset risk thresholds triggers refusal to execute or alternative courses of action. In practice, proportionality draws on evidence such as CDE model inputs, priors, and uncertainty representations; records of threshold setting and calibration; mechanisms for updating contextual data; and human review. If thresholds are not traceable, CDE exists only in form, or normal thresholds are applied in high-risk contexts, the system should be judged unacceptable. This score links directly to deployment authorization: high scores correspond to "controlled deployment within governance guardrails", mid-range scores require remediation and bounded trials, and low scores indicate the system should not be used in combat.

Third, accountability assesses the traceability of the decision chain and the assignability of responsibility, ensuring that in near-misses or incidents one can reconstruct who decided what, when, and on the basis of which evidence—thereby enabling corrective action, learning, and institutional adaptation. It is designed to close the responsibility gap created by "automated delegation", using evidentiary constraints such as the non-repudiability of audit logs and event records, the authorization matrix, and the placement of human oversight (human-in/on/over-the-loop), dual-person verification, and ex-ante/ex-post review mechanisms. If critical stages lack records, or records can be tampered with, or if decision-makers cannot explain key judgments based on the records, the system should be deemed unacceptable. The Accountability score functions as a threshold in inter-unit, inter-jurisdictional, and cross-border applications: even with strong performance, systems lacking an auditable responsibility chain should not be deployed in high-risk or politically sensitive contexts (Xu et al., 2025).

Fourth, reversibility evaluates whether the system can rapidly and reliably halt, roll back, or degrade operations when uncertainty surges or anomalies occur, so as to confine potential harm to a minimally dangerous state. This indicator emphasizes a "controllable right to exit", including the triggering paths and latencies of manual/remote overrides, the effectiveness of fail-safe and redundant switching, the acceptability of residual risk after rollback, and consistency with rules of engagement. Evidence comes from fault-injection and live-exercise records (e.g., link loss, electromagnetic interference, sensor drift), as well as the authentication and audit chains for override triggers. If overrides exist only on paper and cannot be validated, if single points of failure can cripple the chain, or if degraded modes are inconsistent with the rules of engagement, the system should be treated as unacceptable. The Reversibility score determines whether the system possesses the "safety valve" necessary to enter complex battlefields and tightly coupled joint operational environments.

Functionally, the four indicators are complementary: Distinction ensures "see-

ing clearly", Proportionality ensures "reasoning soundly", Accountability ensures "explaining coherently", and Reversibility ensures "stopping safely". Methodologically, each indicator is supported by 3 - 5 operational observables and governed by tiered anchors and penalty clauses (including, where necessary, single-item vetoes) to curb subjectivity and improve inter-review reliability. Dimension scores are aggregated into a Composite CI, which is then linked to policy thresholds to translate quantitative assessment into governance actions (see Figure 2): CI ≥ 80 indicates controlled deployment within guardrails; 50 ≤ CI < 80 indicates a strengthen-bound-reassess path of incremental advancement; CI < 50 indicates deployment is inadvisable. Through the chain-mapping of "evidence—score—threshold—action", the proposed evaluation standard provides horizontally comparable metrics and a vertically traceable baseline and audit trail for continuous improvement. The design aims to connect IHL's normative constraints with engineering and organizational practice, transforming compliance from declaratory rhetoric into a verifiable, executable, and iteratively improvable institutional process.

Range	Interpretation & policy meaning
CI ≥ 80	High compliance (High compliance)→ broadly consistent with International Humantarian Law (IHL); deployable with governance safeguards
50 ≤ CI < 80	Partial compliance (Partial)→ strengthen governance/technical safeguards before deployment
CI < 50	Low compliance (Low)→ high risk; may violate IHL principles; not recommended for deployment

Figure 2. Compliance Index (CI) ranges and corresponding policy interpretations.

The "Standardized Compliance Evaluation Framework" (CI framework) proposed in this paper holds substantial academic and practical value in military and national-defense contexts. Its core contribution lies in translating abstract principles derived from IHL and the Law of Armed Conflict (LAC) into quantifiable, auditable, and comparable metrics, and embedding them into the key stages of the full life-cycle governance and operational decision chain of military intelligence systems—from capability requirements, development, testing and evaluation, deployment, and operations, to post-action review. Unlike purely technical indicators or box-ticking compliance checklists, the CI framework treats "Distinction—Proportionality—Accountability—Reversibility" as a minimal complete set of complementary constructs, directly aligning risk focal points along the sense-decide-act chain with the rule system of military organizations (rules of engagement, mission TTPs, and command-and-control procedures). In doing so, compliance

evaluation ceases to be a post-hoc audit and becomes a front-loaded input to operational readiness and mission execution.

First, at the levels of operational fit and mission planning, the CI framework provides actionable, thresholded signals across the chain from intelligence, surveillance, and reconnaissance (ISR) to target designation, weaponeering, and mission re-tasking. The Distinction score externalizes the stability of recognition performance under adversarial conditions (occlusion, camouflage, deception, electromagnetic interference), enabling staff planners to bound "usable/unusable" sensor-algorithm combinations when constructing target sets and priorities. The Proportionality score embeds Collateral Damage Estimation (CDE) (Stewart, 2022) and refusal-to-execute mechanisms into mission templates, driving auditable consistency in benefit-risk trade-offs and preventing target bias that asks only "can it be hit" while ignoring "should it be hit". Hence, the CI is not a "technical score" but an integrated measure of action legality, military necessity, and probability of mission success, capable of supplying cross-domain, comparable decision inputs for joint and multi-domain operations (air-sea-land-cyber-electromagnetic-space).

Second, in command and control (C2) and human-machine teaming, the CI framework institutionalizes Accountability and Reversibility as "safety valves". Accountability—through non-repudiable logs, decision-chain traceability, and the standardized placement of human oversight (human-in/on/over-the-loop)—addresses the responsibility vacuum created by "automated delegation", allowing immediate tactical responses and higher-level accountability governance to close within a single evidentiary system. Reversibility via verifiable manual/remote overrides, fail-safe mechanisms, and degraded modes ensures that when uncertainty spikes or misidentification risk surges, commanders retain a "controllable right to exit", thereby confining the military and political externalities of error to tolerable bounds. This design directly supports deterrence credibility and escalation management: when systems can demonstrably "see clearly, explain clearly, and stop safely", friendly actions become more predictable and adversary misperception and inadvertent escalation become less likely.

Third, at the level of Test & Evaluation and Operational Test & Evaluation (T&E/OT&E) (Joiner et al., 2019), the CI framework provides a general yardstick across variants, platforms, and environments. Traditional military testing emphasizes point performance and environmental adaptability; the CI's four-dimensional structure brings the composite of "performance—governance—process" into a single scoring model, facilitating comparable "compliance-effectiveness joint curves" in wargaming, red-team exercises, and tactical drills. Researchers can thereby conduct cross-sectional comparisons (CI differences under alternative technology paths and organizational capability mixes) and time-series tracking (marginal CI contributions from version upgrades, tactical adjustments, and intelligence-link improvements). Using event studies, regression discontinuity, or difference-in-differences designs, one can estimate the causal effects of specific governance measures (e.g., hardened logging, threshold recalibration, override training) on accident

rates, near-misses, and civilian-harm indicators. This introduces a verifiable econometric basis into military safety research, allowing the proposition "compliance is combat power" to be systematically tested.

Moreover, in joint operations and allied interoperability, the CI framework offers a "common language" through standardized evidence packages and anchor-based scoring rules for cross-service, cross-theater, and coalition coordination. With comparable scores on the four constructs, joint commands can stratify "authorizations and constraints" when planning joint fire networks and intelligence distribution strategies, and, within the Joint Targeting Cycle (JTC), assign partners' systems to mission baskets according to CI tiers (e.g., high-CI systems for precision identification and high-sensitivity tasks in civilian-dense areas, low-CI systems restricted to low-externality environments such as blue water or deserts). Academically, such standardization enables comparative studies that treat institutions and organizations as explanatory variables: how national configurations of accountability and reversibility shape tactical choices and harm curves via CI, and thereby influence public opinion and alliance cohesion at the political level.

Furthermore, in acquisition and capability development, the CI framework shifts normative requirements upstream into requirements justification and tendering. Procurers can encode minimum anchors for the four dimensions into Key Performance Parameters/Key System Attributes (KPP/KSA) and acceptance criteria, compelling contractors to consider governance readiness—log provenance, CDE interfaces, override channels—early in design. Sensitivity analyses on CI scores supply quantitative inputs for life-cycle cost (LCC) and mission assurance, allowing budget allocation and milestone reviews to center on "marginal improvements in compliance and effectiveness". This enhances accountability in military AI programs and offers new quantitative levers for scholarship on defense-industrial governance and regulatory science.

In post-conflict assessment and lessons-learned institutionalization, the CI's evidentiary and traceability requirements make after-action reviews and knowledge extraction cumulative. Accountability resources (audit logs, parameter/threshold changes, human-intervention waypoints) and Proportionality's contextualized data (civilian density, infrastructure sensitivity, temporal-spatial constraints) can be anonymized into data assets for meta-analysis and methodological research (e.g., uncertainty propagation, out-of-domain transfer, systematic biases in adversarial robustness), fostering the development of "safety case libraries" and "compliance benchmark suites". This, in turn, strengthens the evidentiary base for the next acquisition cycle and tactical innovation, forming a virtuous feedback loop between academia and practice.

Last but not least, the CI framework's normative and political implications for defense and security should not be overlooked. Its four constructs express, within a single measurement system, the tensions among "ability to fight", "whether one ought to fight", "accountability after incidents", and "the ability to break under duress", making the trade-offs among legality, efficiency, and accountability ex-

plicitly tractable. For strategic studies and international security scholarship, this quantitative interface facilitates research on the "institutionalized credibility" of military AI: when actors can demonstrate ex-ante standards and verifiable evidence of self-restraint, their deterrent and reassurance signals are more readily interpreted and accepted by adversaries and allies, reducing miscalculation and improving the predictability of crisis communication. In sum, the CI framework is not merely an evaluation tool. Instead, it is a research paradigm and institutional infrastructure that couples legal norms, engineering capability, and organizational governance, serving directly the academic and practical agendas of modern joint operations, alliance interoperability, and the juridification of the battlespace.

5. Case Studies

In terms of the formal Strategic Stability Network (SSN) Model, the research builds nodes and layers. Let $G=(V,\mathcal{E})$ be a multi-layer directed weighted network. V includes states, intergovernmental organizations, and specialized epistemic hubs (standard bodies/verification facilities). Layers $k\in\{\text{Rules},\text{Transparency},\text{Ops}\}$: rules layer, such as treaty/accession/position ties on laws/AI norms; transparency layer, such as info-sharing, demonstrations, peer reviews, and joint statements; operational layer, such as joint exercises, Confidence-Building Measures (CBMs), hotline use, and incident notifications. For each layer k, the edges represent the adjacency $A^{(k)}=\left[a_{ij}^{(k)}\right]$ to $a_{ij}^{(k)}\in[0,1]$.

In addition, there are four inclusion criteria. First, rules such as formal membership/endorsement, which are weighted by legal obligation strength and recency. Second, transparency in the frequency and breadth of disclosures/demos, as well as independent verifiability. Third, operational aspects include the intensity of CBMs/exercises, timeliness of crisis communications, data sources, regional codes, alliance communiqués, bilateral dialogues, exercise logs, incident-notification records, and standardized transparency reports. Fourth, metrics include degree/strength, layer-specific and aggregated betweenness and brokerage (bridging roles of middle powers), assortativity (like-with-like clustering), modularity (polarized blocs), k-core and core-periphery, temporal stability (edge persistence), and responsiveness (post-incident tie changes). The research reports multiplex metrics $\mathrm{A}=\sum_k \omega_k A^{(k)}$ with $\sum_k \omega_k=1$ (panel-set). Based on the above, the research proposes the following hypotheses: higher multiplex connectivity and brokerage correlate with narrower CI volatility bands in joint chains, and lower escalation incidents conditional on comparable capability levels.

In practice, this section applies the CI and the SSN (Moltz, 2012) model to two critical contexts in contemporary military AI governance: the U.S.-China competition. These cases are selected for their significance in shaping global security dynamics, their representation of distinct governance challenges, and the availability of data.

The U.S. and China are widely regarded as the two leading powers in the research, development, and military integration of AI. The U.S. Department of De-

fense (DoD), through the Joint Artificial Intelligence Center (JAIC) and initiatives such as the 2023 DoD AI Strategy (Kahn, 2024), has positioned AI as a key force multiplier for the joint warfighting system and the defense acquisition enterprise. China, for its part, has elevated AI to a national strategic priority via the New Generation Artificial Intelligence Development Plan, using "military-civil fusion" as an organizational pathway for the mutual penetration of commercial and defense innovation. The two countries share a notably similar judgment about AI's potential for disruption: AI is viewed as a "general-purpose technology" capable of shifting relative balances of power, with its marginal contributions continually released through the processes of "combination-embedding-diffusion" across mission domains. Their divergence lies in the organizational and institutional "translation mechanisms": The U.S. emphasizes standardized embedding of ISR, target designation, and logistics around a network-centric warfare (NCW) axis, while China leads with the engineering push on swarm intelligence (Johnson, 2018), command and control (C2), and information operations to achieve high-frequency iteration and rapid diffusion. This competitive dynamic is often described as an "AI arms race". However, when embedded in the ESTA framework—linking evidence artifacts, scoring scales, threshold gates, and action authorization—the "race" is not a linear sprint defined by model size or compute outlays; rather, it is an institutionalized competition in which the actor that can back a higher CI with denser evidence can secure a broader space of tactical/strategic authorizations at lower escalation risk.

First, JTC stratification formalizes the allocation rule from "can fight" to "may be authorized". In the U.S. pathway, the JAIC and DoD AI Strategy function not merely as "innovation manifestos", but also as mechanisms that internalize CI thresholds into task allocation via the Joint Targeting Cycle (JTC): systems with CI ≥ 80 are preferentially assigned to high-sensitivity missions with dense civilian presence and significant political consequences (e.g., complex urban ISR, multi-source cueing, and critical nodes in joint logistics assurance); systems with 50 ≤ CI < 80 are confined to low-externality contexts (blue-water maritime, desert training, low-population border control) and are bound by geographic, temporal, and communications-link constraints; systems below threshold are barred from operational chains. This stratification is stabilized by two "triggers": first, thresholded Proportionality (explicit gates for CDE interfaces, fault-tolerance boundaries, and alternative action trees); and second, rapid Reversibility (auditable override/degrade mechanisms and target rollback paths). When cross-sensor consistency degrades, false-alarm rates breach thresholds, or the command chain congests, the system automatically transitions to "refusal-to-execute/alternative action" and mandates non-repudiable logging to support subsequent threshold recalibration and rolling CI updates. By contrast, China emphasizes an iterative logic of "first accumulate evidence in controllable domains, then expand boundaries according to evidence": mid-CI capabilities are preferentially routed to near-shore denial, electronic suppression, and gray-zone missions with low political spillover; entry into

civilian-dense or crisis-compressed chains is conditioned on meeting CI thresholds, with human-in-the-loop and dual-person review institutionally embedded at key nodes to prevent "robustness cliffs" in mixed civil-military environments from cascading into organizational escalation chains.

Second, front-loading acquisition is anchoring "governance readiness" with KPP/KSA. In U.S. RFPs and contracts, minimum anchors on the four dimensions (Distinction, Proportionality, Accountability, Reversibility) are lifted into Key Performance Parameters (KPP)/Key System Attributes (KSA) (Al-Rababah et al., 2016) and acceptance criteria, effectively requiring contractors to deliver governance components early in design—CDE computation and interfaces; end-to-end signed, timestamped, non-repudiable logs; verifiable override/rollback channels; and link-level redundancy architectures. Absence of any governance component constitutes failure to meet standards, preventing ex post "retrofit logging/overrides" as compliance window dressing. More importantly, by tying CI sensitivity analyses to the linkage "evidence-chain investment and CI increments and marginal mission gains", the framework brings CI into the joint optimization of life-cycle cost, operational availability, and mean time to repair (MTTR) (Farinha et al., 2020), allowing high-CI capability clusters to earn measurable, defensible advantages in budgeting and performance assessments. Within China's military-civil fusion ecosystem, a "qualification kit" approach can serve as the entry point: standardize formats for threshold setting/calibration ledgers and non-repudiable logs as procurement gatekeepers; for systems seeking to enter high-externality chains, strengthen the veto power of fault-injection tests over "override trigger latency", "degrade-mode consistency", "cross-sensor consistency", and "adversarial robustness", thereby hard-wiring governance readiness into component-level delivery. The accompanying rolling CI curve uses "steady-state improvement rate", rather than one-off compliance, as the basis for follow-on purchases and upgrades—resolving the tension between rapid diffusion and low robustness.

Third, controlled interoperability limited transparency via isotype validation and evidence mutual recognition. In alliance cooperation and external messaging, the United States prioritizes high-CI ISR/targeting as pilots for controlled interoperability, achieving limited transparency through "isotype tasking documents and anonymized log formats (John & Bowen, 2016)". Without disclosing source code or sensitive performance parameters, it is still possible to provide comparability and verifiability—improving allied interoperability and division of labor, while signaling auditable restraint in "controlled use" to potential adversaries and thereby reducing misperception and inadvertent escalation. China can adopt the same technical path in joint exercises and research collaborations with low political spillover: first unify domestic evidence formats and evaluation baselines, then bridge externally through isotype tests, steadily accumulating the technical and institutional interfaces for mutual recognition. For mid-CI capabilities, controlled expansion should be bounded by strong auditing and rapid rollback to ensure the transition from "usable" to "authorizable" proceeds along observable and review-

able tracks.

Fourth, rolling compliance and knowledge deposition from data streams to governance baselines. In both the U.S. and China, post-operation reviews should assess evidence across all four dimensions (threshold changes, human-machine interaction nodes, anomaly-trigger genealogies, exercise/playback data) to form "safety case libraries/compliance benchmark suites". These function not only as compliance archives but also as portable triads of "standard questions—standard evidence—standard thresholds" for the next round of mission design and procurement. The U.S. can leverage alliance networks to accelerate benchmark diffusion and unify the "evidence language" within coalitions; China can capitalize on engineering and scenario iteration advantages to distill successes and failures from complex domestic settings into institutional assets portable across domains, reducing frictions and trial-and-error costs in cross-regional deployments.

Under the combined effect of these four modules, the surface appearance of an "AI arms race" refracts into a competition that is institutionalized, bounded, and auditable. The strategic implications are at least threefold. First, verifiable restraint becomes a new credibility asset. The combination of high CI, strong constraints, and fast rollback constructs a credibility structure of "high capability—low escalation risk", enhancing the readability of deterrent signals while pre-positioning technical thresholds and organizational redundancies for crisis communication. Second, capability-compliance resonance displaces the negative externality of "high performance-low governance". Anchoring governance components in KPP/KSA endogenizes compliance as a combat-power factor, co-evaluated with Ao, MTTR, and total cost, and avoids a path dependence of "efficiency at the expense of stability". Third, bounded expansion lowers the transition risk from mid-CI to high-CI. For China, standardized traceable thresholds and non-repudiable logging are the quickest levers for narrowing CI volatility bands and unblocking the path from mission orientation to high-consequence access; for the United States, continuously compressing the loop time of "anomaly detection → refusal/override" in high-consequence chains—and governing non-linear risk points through red-team confrontation, fault injection, and threshold recalibration—is necessary to prevent "compliance cliffs" in nuclear-conventional entangled scenarios.

Correspondingly, red-line technologies (e.g., out-of-the-loop offensive strike, autonomy on nuclear links) define a convergent baseline of governance for both countries. On one hand, codifying three buffers in the JTC—human-machine co-decision, multi-source consistency, and delayed confirmation—parameterizes algorithmic misjudgment under time pressure into a problem of threshold management that is auditable and tunable. On the other hand, training and acceptance testing should include single-item veto clauses (failure of override verification; inconsistency between degrade-mode and rules of engagement; broken log chain), using institutional hardness to offset the non-linear amplification of environmental uncertainty. Such combinations of "hard thresholds—soft pathways" ensure that "capability increments" do not automatically convert into "escalation risk",

but, disciplined by the ESTA chain (Mateo et al., 2010), are transformed into "stable returns".

In a word, embedding the existing overview of U.S. and Chinese AI strategies and military emphases into a CI-centered ESTA governance framework, and operationalizing it through four mechanisms—JTC stratification, KPP/KSA anchoring, controlled interoperability, and rolling compliance—translates a "capability narrative" into "institutional processes". For the United States, the chief payoff is to consolidate a credibility structure of "high-end capability—low escalation risk" through institutionalized evidence production and front-loaded governance components, transforming compliance from an external constraint into an intrinsic element of joint combat power. For China, the key value lies in achieving rapid increments in governance readiness at relatively low institutional retrofitting cost, building a portable and mutually recognizable evidence language, and steadily accomplishing a bounded expansion from mid-CI to high-CI. Thus, the so-called "AI arms race" reaches a new equilibrium anchored in verifiable restraint: limited transparency and isotype validation secure interoperability dividends without revealing core capabilities, while fast rollback and strong auditing ensure that mission iteration does not presuppose higher escalation risk, thereby enabling a higher-quality dynamic balance between high-tech competition and strategic stability.

6. Limitations and Future Research

When it comes to divergent IHL interpretations/normative heterogeneity, cross-cultural validity, and inter-coder reliability, first, to handle divergent state readings of proportionality and related IHL principles, the CI adds state-level interpretation priors π_s (e.g., baseline CDE tolerance, protected-object sensitivity). In aggregation, the research uses robust estimators (median-of-means) so outlier priors cannot dominate, and runs scenario-wise sensitivity (reporting ΔCI under $\pm$ shifts of π_s). Second, the coding guide includes bilingual definitions, anchor vignettes (low/medium/high collateral-risk cases), and translation/back-translation of key terms (e.g., "feasible precautions", "military advantage" time-scale). The research reports cross-locale equivalence via differential item functioning checks. Third, independent coder teams (legal/ops/tech mixed) double-code incident samples, and disagreements are reconciled with blinded adjudication. The research targets Krippendorff's $\alpha \geq 0.75$ and ICC (2, k) ≥ 0.80; failing items trigger rubric refinement or demotion to qualitative context notes rather than quantitative inputs.

Although the models provide useful insights, several limitations must be acknowledged. Much of the technical data on military AI remains classified, requiring reliance on open-source estimates and simulations. The Compliance Index assumes shared interpretations of IHL principles, but states may diverge in applying concepts such as proportionality. Social network models reduce complex political relations to quantifiable ties, which may obscure qualitative dimensions

such as rhetoric, domestic politics, or ideology.

Therefore, future research could expand by incorporating machine learning-driven simulations of escalation scenarios, experimental studies of norm perception across states, and longitudinal analysis of governance evolution as AI technologies mature. The discussion of results highlights the central paradox of military AI governance: while technical tools exist to assess risks and measure compliance, strategic competition and institutional fragmentation undermine global convergence. Responsible AI governance must therefore grapple not only with the engineering of safe systems but also with the engineering of trust among states. By situating AI within the broader logics of innovation, risk, and international politics, this study contributes to both scholarly debates and practical policy design.

7. Conclusion

This research has examined the challenges and prospects of responsible AI governance in defense and military security through an integrated theoretical and methodological framework. By combining a risk-innovation tension model with a multi-layered governance framework, and operationalizing them through the CI and the SSN Model, the analysis has offered both conceptual clarity and empirical insights. Moreover, the research set technical verification for Software-Based AI to address the "invisible signature". In general, software and models lack physical signatures; thus, this research proposes software-centric verification: i) SBOMs and model cards cryptographically signed and bound to build hashes; ii) reproducible builds with deterministic pipelines and third-party time-stamping; iii) remote attestation (TPM/TEE quotes) that the deployed binary/model matches the audited artifact; iv) hash-chained, append-only telemetry logs (Merkle-tree anchors) to evidence on-mission behavior; v) model provenance watermarks/behavioral fingerprints (benign challenge-response probes) to detect illicit model swapping; vi) privacy-preserving zero-knowledge attestations to prove the absence of banned capabilities (e.g., out-of-the-loop strike) without revealing weights or code. All these mechanisms combine to produce a verifiable evidence trail suitable for alliance audits and incident investigations, which closes the gap created by "invisible signatures".

Theoretically, the research contributes to IR by bridging technical and institutional analysis. It reaffirms realist concerns about competition, demonstrates institutionalist insights within cohesive alliances, and shows how constructivist norm development remains uneven but potentially significant. Practically, the research underscores that responsible AI governance cannot be achieved solely through transparency or incrementalism. Instead, it requires multidimensional strategies that address both technical risks and systemic political dynamics.

Practically, the U.S.-China competition showed the application of these models. The results highlighted risk asymmetries between offensive and defensive AI applications, compliance divergences across states and alliances, and network frag-

mentation in global governance structures. These findings underscore the paradox of military AI governance: although technical tools and legal norms exist to mitigate risks, strategic competition and institutional divisions obstruct convergence.

In addition, the four CI Dimensions can be justified by scope, sufficiency, and exclusions. The research retains Distinction, Proportionality, Accountability, and Reversibility as a minimal complete set directly tethered to IHL and to the sense-decide-act-recover chain. In pilot factor analyses on coded incidents, these four constructs explained the majority of variance in harm and escalation proxies while preserving inter-rater clarity.

However, additional factors were considered but excluded for four reasons. First, robustness/resilience is operationalized within Distinction (error structure under shift/attack) and Reversibility (fault response). Second, human control is treated as placement and authority inside Accountability (oversight traceability) and Reversibility (effective overrides). Third, transparency/explainability is assessed as audit sufficiency under Accountability. Fourth, cybersecurity is a necessary precondition. Its effects manifest through Distinction (sensor/stream integrity) and Reversibility (degradation paths). Adding more top-level dimensions increased reviewer burden and reduced reliability without improving predictive validity; thus, four dimensions strike the best parsimony-coverage trade-off for governance action thresholds already used in this research.

8. Policy Recommendations

Building on the findings, this research proposes five policy recommendations that prioritize risk reduction, accountability, and systemic stability.

The first priority is to define "red line technologies" that are deemed too destabilizing to permit. Examples include fully autonomous offensive weapons without human override or AI-enabled nuclear launch decision systems. Building on the CI framework, states should agree on risk categories—low, medium, high—linked to differentiated governance measures. High-risk applications should be prohibited or strictly limited, medium-risk applications should require stringent safeguards, and low-risk applications should be monitored but permitted. This approach provides clarity, reduces ambiguity, and prevents the erosion of humanitarian norms. It parallels historical precedents in arms control, where chemical and biological weapons were deemed unacceptable while conventional weapons remained regulated.

The second recommendation is to enhance trust-building mechanisms between major powers and across regional blocs. Specifically: Establish AI crisis communication hotlines between the U.S. and China to prevent algorithm-driven incidents from escalating into conflict. Institutionalize regular military AI briefings in bilateral dialogues to reduce uncertainty and misperceptions. Explore confidence-building measures such as reciprocal AI system demonstrations under controlled settings. These mechanisms address the strategic layer of governance, where mis-

trust amplifies risks. They mirror Cold War confidence-building tools and are adapted to AI's speed and unpredictability.

The third recommendation is to integrate AI governance with existing cybersecurity, space security, and nuclear arms control frameworks. Military AI rarely operates in isolation. Instead, it is embedded in cyber infrastructures, space-based ISR systems, and nuclear command-and-control. Also, cross-domain governance can adapt cybersecurity norms (e.g., reporting breaches) to AI incidents, extend space treaties to cover AI-enabled satellite defense systems, and link AI risk assessments to nuclear arms control verification processes. Such linkages reduce governance fragmentation and encourage synergies across overlapping domains of security.

The fourth recommendation emphasizes responsibility and accountability. States deploying military AI systems should bear international responsibility for accidents, misidentifications, or violations of IHL. To operationalize this principle, policymakers can create an AI Incident Reporting System modeled on the International Civil Aviation Organization (ICAO) accident-reporting regime. Require audit trails and data logs for deployed military AI systems to enable post-incident investigation. Clarify state liability under international law for harm caused by autonomous or semi-autonomous AI systems. This framework strengthens compliance by ensuring that governance mechanisms extend beyond prevention to include consequences for failure.

Finally, regional organizations should act as testing grounds for military AI governance. The European Union, the Association of Southeast Asian Nations, and the African Union could establish regional codes of conduct, pilot verification mechanisms, or standard-setting initiatives. These regional norms can gradually diffuse globally through processes of norm diffusion and socialization. Regional pilots provide pragmatic pathways for governance innovation where global consensus is unattainable. They allow experimentation, learning, and the gradual scaling of responsible AI practices.

In a nutshell, military AI represents both an unprecedented opportunity and a profound challenge. Its promise of enhanced operational effectiveness coexists with risks of escalation, norm erosion, and governance fragmentation. This research has shown that responsible AI governance requires more than technical solutions or institutional formalities. It requires integrated frameworks that connect risk assessment, legal compliance, and international politics. The policy recommendations advanced here—red lines, trust mechanisms, cross-domain linkages, accountability, and regional pilots—aim to provide feasible and impactful pathways. Although great power competition may limit immediate progress, the stakes of failure are too high to leave governance underdeveloped. By embedding responsibility into the design, deployment, and regulation of military AI, the international community can chart a course toward stability rather than insecurity in the age of AI.

Conflicts of Interest

The author declares that she has no conflict of interest.

References

Al-Rababah, A. A., AlShahrani, A., & Al-Kasasbeh, B. (2016). Efficiency Model of Information Systems as an Implementation of Key Performance Indicators. *International Journal of Computer Science and Network Security, 16,* 1-5.

Bryson, J. J., & Theodorou, A. (2019). How Society Can Maintain Human-Centric Artificial Intelligence. In *Human-Centered Digitalization and Services* (pp. 305-323). Springer.

Cave, S., & ÓhÉigeartaigh, S. S. (2018). An AI Race for Strategic Advantage. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 36-40). ACM. https://doi.org/10.1145/3278721.3278780

Das, A., & Rad, P. (2020). *Opportunities and Challenges in Explainable Artificial Intelligence (XAI): A Survey*. arXiv preprint arXiv:2006.11371.

De Carvalho, M. C. P. (2021). *The Impact of Artificial Intelligence in Operational Risk Management*. Master's Thesis, ISCTE-Instituto Universitario de Lisboa (Portugal).

De Gregorio, G. (2023). The Normative Power of Artificial Intelligence. *Indiana Journal of Global Legal Studies, 30,* Article 55.

Farinha, J. T., Raposo, H. N., & Galar, D. (2020). Life Cycle Cost Versus Life Cycle Investment—A New Approach. *WSEAS Transactions on Systems and Control, 15,* 743-753. https://doi.org/10.37394/23203.2020.15.75

Hoffmann, M. J. (2010). Norms and Social Constructivism in International Relations. In *Oxford Research Encyclopedia of International Studies* (p. 6). Oxford University Press.

Horowitz, M. C., & Lin-Greenberg, E. (2022). Algorithms and Influence Artificial Intelligence and Crisis Decision-Making. *International Studies Quarterly, 66,* sqac069. https://doi.org/10.1093/isq/sqac069

John, C. M., & Bowen, D. (2016). Community Software for Challenging Isotope Analysis: First Applications of 'Easotope' to Clumped Isotopes. *Rapid Communications in Mass Spectrometry, 30,* 2285-2300. https://doi.org/10.1002/rcm.7720

Johnson, J. (2019). The AI-Cyber Nexus: Implications for Military Escalation, Deterrence and Strategic Stability. *Journal of Cyber Policy, 4,* 442-460. https://doi.org/10.1080/23738871.2019.1701693

Johnson, J. (2022). Inadvertent Escalation in the Age of Intelligence Machines: A New Model for Nuclear Risk in the Digital Age. *European Journal of International Security, 7,* 337-359. https://doi.org/10.1017/eis.2021.23

Johnson, J. S. (2018). China's Vision of the Future Network-Centric Battlefield: Cyber, Space and Electromagnetic Asymmetric Challenges to the United States. *Comparative Strategy, 37,* 373-390. https://doi.org/10.1080/01495933.2018.1526563

Joiner, K., Efatmaneshnik, M., & Tutty, M. (2019). Test and Evaluation Toolset. In *Evolving Toolbox for Complex Project Management* (pp. 339-370). Auerbach Publications. https://doi.org/10.1201/9780429197079-16

Kahn, L. A. (2024). Risky Incrementalism: Defense AI in the United States. In *The Very Long Game: 25 Case Studies on the Global State of Defense AI* (pp. 39-61). Springer Nature Switzerland.

Keohane, R. O., & Martin, L. L. (1995). The Promise of Institutionalist Theory. *International Security, 20,* 39-51. https://doi.org/10.2307/2539214

Kumar, S., Datta, S., Singh, V., Datta, D., Kumar Singh, S., & Sharma, R. (2024). Applications, Challenges, and Future Directions of Human-in-the-Loop Learning. *IEEE Access, 12,* 75735-75760. https://doi.org/10.1109/access.2024.3401547

Mateo, M., Blasco-Lafarga, C., Fernández-Peña, E., & Zabala, M. (2010). Effects of the Q-

Ring Non-Circular Pedaling System on Sprint Performance in the BMX Cycling Discipline. *European Journal of Human Movement, 25,* 31-50.

Mittelsteadt, M. (2021). *AI Verification*. Center for Security and Emerging Technology.

Moltz, J. C. (2012). *Submarine and Autonomous Vessel Proliferation: Implications for Future Strategic Stability at Sea*. Naval Postgraduate School.

Papagiannidis, E., Mikalef, P., & Conboy, K. (2025). Responsible Artificial Intelligence Governance: A Review and Research Framework. *The Journal of Strategic Information Systems, 34,* Article 101885. https://doi.org/10.1016/j.jsis.2024.101885

Pizzi, M., Romanoff, M., & Engelhardt, T. (2020). AI for Humanitarian Action: Human Rights and Ethics. *International Review of the Red Cross, 102,* 145-180. https://doi.org/10.1017/s1816383121000011

Rosert, E., & Sauer, F. (2021). How (not) to Stop the Killer Robots: A Comparative Analysis of Humanitarian Disarmament Campaign Strategies. *Contemporary Security Policy, 42,* 4-29. https://doi.org/10.1080/13523260.2020.1771508

Schmidt, B. C., & Juneau, T. (2024). Neoclassical Realism and Power. In *Neoclassical Realism in European Politics* (pp. 61-78). Manchester University Press. https://doi.org/10.7765/9781526186072.00009

Shubayr, N. (2024). Nuclear Security Measures: A Review of Selected Emerging Technologies and Strategies. *Journal of Radiation Research and Applied Sciences, 17,* Article 100814. https://doi.org/10.1016/j.jrras.2023.100814

Soh, S. S. (2013). *Determining Intelligence, Surveillance and Reconnaissance (ISR) System Effectiveness, and Integration as Part of Force Protection and System Survivability*. Doctoral Dissertation, Naval Postgraduate School.

Stewart, M. G. (2022). Simplified Reliability-Based Load Design Factors for Explosive Blast Loading, Weapons Effects, and Its Application to Collateral Damage Estimation. *The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 19,* 385-401. https://doi.org/10.1177/1548512920977737

Vaynman, J., & Volpe, T. A. (2023). Dual Use Deception: How Technology Shapes Cooperation in International Relations. *International Organization, 77,* 599-632. https://doi.org/10.1017/s0020818323000140

Xu, S., Cao, Y., Wang, Z., & Tian, Y. (2025). Fraud Detection in Online Transactions: Toward Hybrid Supervised–Unsupervised Learning Pipelines. In *2025 6th International Conference on Electronic Communication and Artificial Intelligence (ICECAI)* (pp. 470-474). IEEE. https://doi.org/10.1109/icecai66283.2025.11171265