Journal of Intelligent Decision Support System (IDSS) 9 (2) (2026), pp. 138-147

Published by: Institute of Computer Science (IOCS)

Journal of Intelligent Decision Support System (IDSS)

Journal homepage: www.idss.iocspublisher.org

Escalation-Aware Governance Framework for Autonomous Warfare Decision Support Systems

Ni Luh Meliana Liberty[1], H.A.Danang Rimbawa[2], Bambang Suhardjo[3]

[1,2,3] *Department of Cyber Defense Technology, Republic of Indonesia Defense University, Bogor, Indonesia*

Article Info

Article history:

Received May 30, 2026
Revised Jun 18, 2026
Accepted Jun 30, 2026

Keywords:

AI governance;
Autonomous warfare;
Decision support systems;
Human oversight;
Strategic stability.

ABSTRACT

Artificial intelligence increasingly transforms military decision-making through autonomous targeting, predictive battlefield analytics, cyber operations, and algorithm-assisted command systems. Although these technologies improve operational speed, they create governance risks when machine-speed decisions exceed human oversight, legal accountability, and escalation control. This study analyzes governance asymmetry in AI-driven autonomous warfare decision support systems and develops an escalation-aware governance framework. A structured qualitative review using transparent search procedures and thematic synthesis was conducted across 33 peer-reviewed and institutional sources on military AI, autonomous weapons, cybersecurity, and AI governance. The synthesis reveals three interconnected governance failures governance asymmetry, symbolic human control, and escalation compression that reinforce one another by weakening accountability, reducing substantive human judgment, and increasing strategic instability. The study proposes a human-centered escalation-aware governance framework integrating eight operational control layers: human authorization checkpoints, explainability, auditability, cybersecurity resilience, cognitive security, escalation-control mechanisms, emergency override, and post-action review. The framework contributes to intelligent decision support literature by translating autonomous warfare risks into embedded governance layers for preserving accountability and strategic stability.

Corresponding Author:

Ni Luh Meliana Liberty,
Master Programee of Cyber Defense's Department,
Republic of Indonesia Defense University, Bogor, Indonesia
Kawasan IPSC Sentul, Jl. Anyar, Sukahati, Kec. Citeureup, Kabupaten Bogor, Jawa Barat 16810
Email: luh.liberty@tp.idu.ac.id

INTRODUCTION

Artificial intelligence has changed the logic of military decision-making by transforming battlefield operations from human-paced command processes into machine-speed decision environments. Contemporary military systems increasingly integrate autonomous targeting, predictive intelligence, cyber defence, drone coordination, sensor fusion, and algorithm-assisted command-and-control functions. This transformation does not merely add new tools to existing military institutions; it alters the tempo, structure, and accountability of military decisions. AI-enabled systems can process large-scale battlefield data and generate operational recommendations faster than conventional human deliberation, thereby creating a strategic environment in which speed becomes both a military advantage and a governance vulnerability (Horowitz & Kahn, 2024; Johnson, 2021; Payne, 2021)

In the context of intelligent decision support systems, the governance problem does not arise only from autonomous weapon execution but from the way AI systems structure, rank, filter, and frame military options for human commanders. AI-based decision support systems may shape what commanders perceive as urgent, credible, targetable, or proportionate before any weapon is launched. Their influence therefore occurs upstream of the use of force, in the production of situational awareness, threat prioritization, target recommendation, and response selection. This makes autonomous warfare decision support systems a high risk socio technical architecture rather than a neutral computational tool (Dorton & Harper, 2022; Johnson, 2023; Macrae, 2022)

The central governance problem is that autonomous warfare decision support systems develop faster than the legal, ethical, institutional, and operational mechanisms required to control them. International debates on lethal autonomous weapon systems, responsible military AI, and meaningful human control show that states increasingly acknowledge the risks of autonomy in the use of force, yet global governance remains fragmented and largely dependent on soft law principles, national doctrines, and voluntary political commitments (Bode et al., 2023; Cross, 2021; State, 2023). The North Atlantic Treaty Organization's (NATO) revised AI strategy emphasizes responsible use, interoperability, and protection against adversarial uses of AI, including AI-enabled information operations and disinformation (Organization, 2024). However, principles alone do not ensure accountability when autonomous systems operate under time pressure, uncertainty, and adversarial manipulation.

This mismatch creates governance asymmetry, defined in this study as the structural gap between machine-speed military decision-making and human-centered institutional oversight. Governance asymmetry occurs when autonomous systems accelerate the observe orient decide act (OODA) cycle while regulatory, command, and accountability mechanisms remain designed for slower human-paced warfare (Boyd, 1996; Johnson, 2023) .The risk is not only that machines may make errors, but that human institutions may lose the practical capacity to understand, challenge, delay, or override AI-generated military recommendations in time-sensitive environments. This condition is particularly relevant to intelligent decision support systems because decision support does not remain neutral when its outputs reshape command authority, operational confidence, and escalation pathways (Johnson, 2022; Nadibaidze & Miotto, 2023).

Meaningful human control is therefore a central issue. Existing literature emphasizes that humans should retain authority over decisions involving the use of force because machines lack moral agency, contextual judgment, and legal responsibility (Amoroso & Tamburrini, 2023; Ferl, 2024). Nevertheless, the presence of a human operator does not necessarily guarantee meaningful control. In machine-speed conflict, a human may remain formally in the loop but practically function only as a procedural validator of algorithmic recommendations. Automation bias, information overload, black-box models, and compressed decision windows can transform human oversight into symbolic human control, where accountability remains formally human but operational judgment becomes algorithmically dependent (Christie et al., 2023; Horowitz & Kahn, 2024).

A related and equally critical dimension is escalation compression, defined here as the reduction of time available for verification, diplomatic communication, and command review due to autonomous or semi-autonomous decision cycles. When AI-driven systems accelerate the tempo of threat classification and response planning, the windows for human deliberation, strategic restraint, and diplomatic de-escalation are correspondingly narrowed. This compression is especially dangerous because it can trigger disproportionate or unintended military responses before human institutions can intervene, thereby intensifying strategic instability (Johnson, 2022, 2023; Nadibaidze & Miotto, 2023). Governance asymmetry, symbolic human control, and escalation compression are thus analytically distinct yet mutually reinforcing: each deepens the conditions that make the others more likely.

Existing studies have examined autonomous weapons ethics, international humanitarian law, strategic stability, cybersecurity, and AI governance separately (Boutin, 2023; Rosert & Sauer, 2021). However, fewer studies integrate these dimensions into a decision-support governance framework that explains how AI-driven military systems generate interconnected governance failures and how those failures can be translated into operational governance mechanisms. This study addresses that

gap by investigating how autonomous warfare decision support systems generate governance asymmetry, how machine-speed decision-making transforms meaningful human control into symbolic oversight, and what governance framework can preserve accountability, human oversight, and escalation control.

The contribution of this study lies in integrating governance asymmetry, symbolic human control, and escalation compression into a single operational architecture and translating abstract AI governance principles into eight concrete, embedded control layers an escalation-aware governance framework.

METHOD

This study employed a structured qualitative review using transparent search procedures and thematic synthesis. This formulation is used deliberately because the article does not claim to conduct statistical meta-analysis or empirical battlefield testing. Instead, it develops a conceptual governance framework from structured literature identification, transparent selection criteria, and cross-source thematic synthesis. The approach is appropriate for examining autonomous warfare decision support systems because the topic involves legal, ethical, strategic, operational, and socio-technical dimensions that cannot be captured through a single disciplinary method. The coding logic was informed by thematic analysis principles and literature review guidance, while the screening process followed transparency standards for systematized reviews rather than a full bibliometric meta-analysis protocol (Braun & Clarke, 2021; Glaser & Strauss, 2024; Page et al., 2021; Snyder, 2023; Timonen & Conlon, 2023).

The literature identification was designed as a structured and reproducible search strategy covering publisher platforms and academic databases. Search strings combined military AI, autonomous weapons, decision support, human oversight, explainability, accountability, and escalation terms. The main search strings included "autonomous weapons systems" AND "meaningful human control", "military AI governance" AND accountability, "responsible military AI" AND "human oversight", "AI decision support systems" AND military, "AI-DSS" AND targeting, "military decision support" AND artificial intelligence, "command-and-control" AND AI AND decision-making, "machine-speed warfare" AND escalation, "AI" AND "strategic stability", "inadvertent escalation" AND "intelligent machines", and "explainability" AND "traceability" AND "autonomous weapons". The phrase "autonomous warfare" AND "decision support" was treated only as a supplementary query because preliminary source checking showed that it retrieves more policy and professional commentary than peer-reviewed decision-support literature.

To avoid overstating the procedure as a full bibliometric or PRISMA-based systematic review, this study does not report database-export counts that cannot be independently reproduced in the manuscript. Instead, the review reports the final analytical corpus used for thematic synthesis. These sources were retained because they directly addressed at least one of the following criteria: military AI governance, autonomous warfare, decision-support relevance, human oversight, explainability, traceability, accountability, cybersecurity resilience, cognitive security, or escalation risk. The selection criteria are summarized in Table 1.

Table 1. Literature selection criteria

Criterion Type	Criterion	Description
Inclusion	Topic relevance	Studies discussing military AI, autonomous warfare, autonomous weapons, AI-enabled command systems, or decision support systems
Inclusion	Governance relevance	Studies addressing human oversight, accountability, explainability, traceability, regulation, legal responsibility, ethics, escalation, or cybersecurity
Inclusion	Decision-support relevance	Studies explaining how AI structures, filters, ranks, recommends, or supports decisions in high-risk military/security environments
Inclusion	Source quality	Peer-reviewed journals, reputable academic publishers, and selected institutional governance reports
Inclusion	Time relevance	Mainly 2019–2025, with older sources used only for theoretical or methodological grounding
Exclusion	Civilian AI only	Studies unrelated to military, security, or high-risk decision support contexts
Exclusion	Pure engineering	Technical studies without governance, accountability, oversight, or decision-

focus	making implications

The analysis followed thematic synthesis logic. Open coding identified recurring concepts such as accountability gap, automation bias, explainability, escalation risk, cyber vulnerability, cognitive manipulation, auditability, and human oversight. Axial coding connected these concepts into broader governance mechanisms, especially governance asymmetry, symbolic human control, and escalation compression. Selective coding then mapped each governance failure to possible control mechanisms derived from the literature, producing the proposed escalation-aware governance framework. Analytical validity was strengthened through source triangulation among peer-reviewed studies, institutional principles, and policy frameworks such as DoD Directive 3000.09, the ICRC position on autonomous weapons, NATO's AI strategy, and the Political Declaration on Responsible Military Use of Artificial Intelligence and Autonomy (Cross, 2021; Defense, 2023; Organization, 2024; State, 2023).

Because this review was conducted by a single researcher, inter-coder reliability in the conventional sense (e.g., Cohen's kappa between independent coders) was not calculated. Instead, the following self-audit procedures were applied to ensure analytical consistency: (1) double-pass coding, with a minimum interval of seven days between passes; (2) source triangulation, requiring every proposed control layer to have at least two independent source citations before inclusion; (3) negative-case checking, in which the researcher actively searched for sources contradicting the emergent framework; and (4) a dated audit trail recording all coding decisions, theme revisions, and layer mappings. These procedures are detailed in Appendix B.5.

RESULT AND DISCUSSION

The thematic synthesis indicates that AI-driven autonomous warfare decision support systems generate governance failures through three interconnected mechanisms: governance asymmetry, symbolic human control, and escalation compression. These mechanisms reinforce one another. Machine-speed decision support creates pressure to rely on algorithmic recommendations; reliance on algorithmic recommendations can weaken meaningful human control; weakened human control increases accountability ambiguity and escalation risk. Table 2 summarizes the main thematic findings.

Table 2. Thematic synthesis of governance failures

Theme	Governance Problem	Decision-Support Mechanism	Strategic Impact
Governance asymmetry	Technology develops faster than regulation and institutional oversight	AI accelerates intelligence processing, targeting support, and response planning; distributes operational influence across data pipelines, model design, sensor networks, and procurement choices	Regulatory gap, institutional obsolescence, and distributed accountability
Symbolic human control	Human approval becomes procedural rather than substantive; meaningful control erodes under operational opacity and automation bias	Operators depend on compressed AI outputs under time pressure and uncertainty; black-box models obscure data weighting and inference logic	Accountability erosion, automation bias, and reduced trust generating legal ambiguity
Escalation compression	Decision cycles become faster than diplomatic and command review mechanisms; cognitive vulnerability intensifies misperception	Autonomous systems shorten verification and restraint windows; AI-enabled misinformation and spoofed data contaminate decision inputs	Unintended escalation, crisis instability, and legitimacy erosion

Governance Asymmetry

Governance asymmetry is the foundational failure because it explains why existing control mechanisms become increasingly insufficient. Autonomous warfare decision support systems are designed to increase operational speed, reduce cognitive burden, and improve battlefield responsiveness. Yet these advantages can become governance liabilities when institutions cannot match the speed and complexity of algorithmic decision-making. International governance remains fragmented across different venues, including the UN Convention on Certain Conventional Weapons, REAIM-related initiatives, national AI principles, and voluntary declarations (Bode et al., 2023;

Defense, 2023). This fragmentation means that operational deployment of military AI often advances through national doctrine, procurement, and experimentation before binding international norms become mature.

The governance asymmetry identified in this study is not simply a legal vacuum. It is a socio-technical mismatch between how AI systems operate and how military governance traditionally assigns responsibility. Conventional military governance assumes identifiable command authority, sequential review, human deliberation, and post-action accountability. Autonomous warfare decision support systems complicate these assumptions because they distribute operational influence across data pipelines, model design, sensor networks, operators, commanders, software updates, and institutional procurement choices. The accountability question therefore shifts from "who pressed the button?" to "which human–machine–institutional configuration produced the decision?" This supports recent lifecycle approaches arguing that responsibility and accountability must be established before deployment and maintained across development, testing, operation, and review (Blanchard et al., 2024; Conn & Bode, 2025).

Symbolic Human Control

Meaningful human control becomes difficult when AI systems compress decision time. The literature shows that human control is often treated as a formal requirement, yet formal presence does not guarantee substantive judgment (Amoroso & Tamburrini, 2023; Ferl, 2024). In an ideal human-in-the-loop arrangement, a human operator has the time, information, authority, and competence to assess the AI recommendation before action. In practice, machine-speed warfare can reduce these conditions. Operators may face high volume sensor feeds, probabilistic target classifications, system confidence scores, and tactical urgency. Under these conditions, they may approve AI outputs because rejecting or delaying them appears operationally costly. This is the mechanism through which human control becomes symbolic

Symbolic human control is analytically important because it exposes a weakness in superficial compliance. A system may claim to keep humans in the loop while designing the interface, tempo, and organizational culture in ways that make independent human judgment unrealistic. Automation bias reinforces this condition because users may over-trust AI decision aids, especially when the system appears technically sophisticated or when the operational environment is uncertain (Horowitz & Kahn, 2024). Trust studies in intelligence work also indicate that explainability and performance strongly affect human reliance on AI outputs (Dorton & Harper, 2022). Explainability and traceability are therefore not optional ethical additions; they are preconditions for meaningful oversight. Without explanation, the operator cannot understand why a target is recommended. Without traceability, investigators cannot reconstruct how a decision was produced. Without override authority, human involvement becomes a ritual rather than control (Christie et al., 2023; Floridi, 2023).

Escalation Compression

Escalation compression is the third major finding. It refers to the reduction of time available for verification, restraint, diplomatic communication, and command review due to autonomous or semi-autonomous decision cycles. AI-enabled systems can rapidly classify threats, recommend responses, and coordinate military action. These capabilities may strengthen tactical performance, but they also increase the risk that misclassification, cyber manipulation, spoofed data, or synthetic misinformation triggers disproportionate responses before human institutions intervene. (Johnson, 2022, 2023) argues that AI can intensify command and control risk by accelerating decision cycles while leaving jpjpstrategic interpretation dependent on human cognition. (Nadibaidze & Miotto, 2023) further show that AI's impact on strategic stability is shaped not only by technical capability but also by state perceptions, threat narratives, and distrust. At the highest political level, AI may also affect assessments about whether and how states move toward war, making governance of decision support relevant beyond the tactical battlefield (Erskine & Miller, 2024).

Escalation compression is particularly dangerous in autonomous warfare decision support systems because the decision-support layer shapes what commanders see as urgent, credible, and actionable. When AI-generated outputs structure threat perception, they can narrow the range of

options considered by human decision-makers. A high-confidence AI alert may create pressure to act quickly, even when the underlying data are uncertain or adversarially manipulated. This problem becomes sharper in cyber and cognitive warfare contexts because battlefield information can be contaminated by deepfakes, spoofed sensor data, malicious AI applications, or synthetic narratives (Blauth et al., 2022; Organization, 2024). Autonomous warfare governance must therefore include cognitive security and cyber resilience, not merely weapons control.

The three failures are interconnected through a reinforcing cycle. Governance asymmetry enables the deployment of systems that operate at speeds beyond institutional oversight. These systems produce conditions for symbolic human control, in which operators approve outputs without genuine deliberation. Symbolic human control, in turn, removes the last institutional brake on escalation compression, leaving autonomous decision cycles without meaningful human intervention. Understanding this cycle is essential for designing governance mechanisms that address root causes rather than symptoms.

Proposed Framework: The Human-Centered Escalation-Aware Governance Framework

The proposed framework was derived by mapping each governance failure in Table 2 to a corresponding control mechanism discussed in military AI governance, explainability, accountability, cybersecurity, human agency, and strategic stability literature. This mapping prevents the framework from functioning as a speculative checklist. Human authorization responds to symbolic control; explainability and auditability respond to opacity and accountability gaps, cybersecurity and cognitive verification respond to adversarial manipulation; escalation-control mechanisms respond to machine-speed retaliation; emergency override and post-action review respond to governability and institutional learning. The framework is summarized in Table 3.

Table 3. Human-centered escalation-aware governance framework

Governance Layer	Risk Addressed	Control Mechanism
Human authorization layer	Symbolic human control	Mandatory human review for high-risk targeting, retaliation, and escalation decisions
Explainability layer	Operational opacity	Explanation of recommendation logic, confidence level, uncertainty, and alternative interpretations
Auditability layer	Accountability gap	Human–machine decision logs covering data input, model output, operator action, and command approval
Cybersecurity resilience layer	Adversarial manipulation	Red-teaming, anomaly detection, secure data pipelines, and model validation
Escalation-control layer	Machine-speed retaliation	Delay mechanism, escalation threshold, second-person review, and command-level confirmation
Cognitive security layer	Deepfake and misinformation risk	Source verification, synthetic media detection, and information integrity protocol
Emergency override layer	Loss of governability	Human shutdown authority and fail-safe intervention procedures

The proposed framework reframes autonomous warfare governance from principle declaration into operational architecture. The human authorization layer prevents high-risk decisions from being executed solely through algorithmic momentum. The explainability layer ensures that the system communicates not only what it recommends but why it recommends it, how certain it is, and what uncertainties remain. The auditability layer addresses the responsibility gap by creating a record of human–machine interaction. The cybersecurity layer recognizes that AI governance fails if the input environment can be manipulated. The escalation control layer introduces deliberate friction into machine-speed decision cycles. The cognitive security layer expands governance to perception warfare, where synthetic information can distort operational judgment. The emergency override and post-action review layers preserve governability and institutional learning.

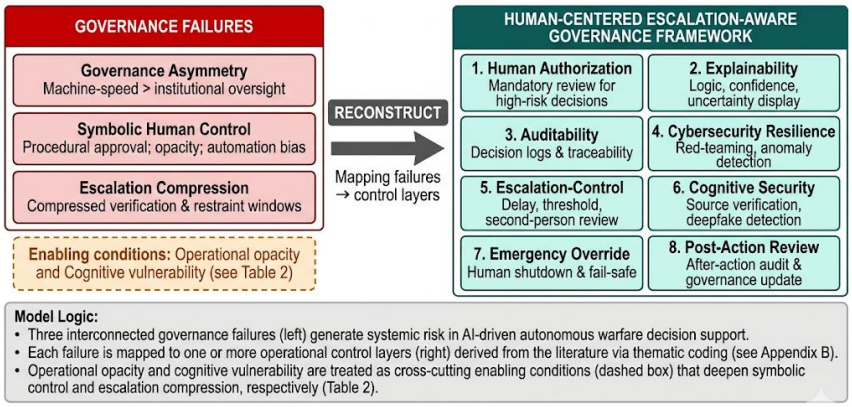

Figure 1. Autonomous warfare governance failure and reconstruction model

Application Scenarios

The framework can be applied as a governance checklist for evaluating autonomous warfare decision support systems before deployment. For example, in an AI-assisted targeting scenario, the system should not merely recommend a target and request approval. It should display uncertainty, identify the data sources supporting the recommendation, flag possible bias or spoofing, require human confirmation for high-risk actions, record the human–machine decision path, and activate escalation delay when the decision may trigger strategic consequences. In this sense, the framework does not reject military AI categorically. It argues that the legitimacy of military AI depends on whether governance mechanisms are embedded into decision-support architecture before deployment rather than improvised after failure (Blanchard et al., 2024; Boshuijzen-van Burken, 2023; Jenkins et al., 2025).

To illustrate how these layers operate in concert, consider an AI-assisted targeting scenario in a contested electronic warfare environment. The system detects an anomalous radar signature and recommends a precision strike. The explainability layer displays a heat map of feature importance, a confidence score of 0.82, and an alternative interpretation suggesting a civilian decoy signature. The cognitive security layer flags that synthetic aperture data from one sensor node failed source-verification protocols. The human authorization layer therefore requires a second-person review by a senior operator before the strike request is elevated to the commander. The escalation-control layer activates a ten-minute deliberation hold because the target is within a de-escalation buffer zone. If the commander approves, the auditability layer logs all inputs, overrides, and timestamps. If the network is compromised mid-process, the emergency override layer allows the commanding officer to revert to manual kill-chain procedures. Following the operation, the post-action review layer maps responsibility across the data pipeline and updates model-validation criteria. This scenario demonstrates that the framework is not merely a checklist but an operational architecture that introduces deliberate friction without paralyzing tactical responsiveness.

A second illustrative scenario involves AI-enabled cyber defence. An autonomous system detects a coordinated intrusion against military communications infrastructure and recommends an immediate retaliatory cyber strike against the presumed origin server. The explainability layer reveals that the attribution confidence is only 0.64 and that the pattern could also match a third-party proxy. The cybersecurity resilience layer reports that the anomaly detection model has not been updated for the latest known adversarial techniques, reducing confidence in the threat classification. The escalation-control layer triggers a mandatory delay because the recommended target resides in the network infrastructure of a nuclear-armed state. The human authorization layer escalates the decision to the strategic command level rather than allowing an automated response. The cognitive security layer cross-references the intrusion pattern against known deception signatures. After the commander decides to pursue defensive containment rather than retaliatory action, the auditability layer records the full deliberation chain, and the post-action review layer schedules a model update. This cyber

scenario demonstrates that the framework applies beyond kinetic targeting to the full spectrum of AI-assisted military decision-making.

Implementation Tensions and Operational Trade-offs

Embedding eight governance layers into machine speed command systems introduces inherent tensions. Mandatory human authorization and escalation delays may reduce the tactical tempo that motivates AI adoption in the first place. Militaries facing time critical threats may perceive explainability requirements as cognitive burdens that slow the observe orient decide act loop. Furthermore, adversarial spoofing sophisticated enough to defeat cognitive security filters may simultaneously trigger false positive emergency overrides, creating denial of service vulnerabilities in the decision chain. These trade offs suggest that governance design must be threat contextual: high risk, high escalation potential decisions warrant full eight layer activation, whereas lower risk support functions may operate with a reduced subset. Future work should develop dynamic layer-activation protocols that balance strategic stability against operational exigency.

CONCLUSION

This study examined AI-driven autonomous warfare decision support systems as a governance problem rather than merely a technological or weapons autonomy issue. Addressing reseacrh question 1, the synthesis shows that autonomous warfare produces governance asymmetry when machine-speed decision-making exceeds the adaptive capacity of legal, institutional, and human oversight mechanisms. This asymmetry weakens accountability because operational influence becomes distributed across algorithms, data, operators, commanders, and institutions.

Addressing reseacrh question 2, the study found that meaningful human control can become symbolic human control when operators formally approve AI recommendations without sufficient time, explanation, authority, or capacity to challenge them. The concept of symbolic human control captures the gap between the formal presence of a human in the decision loop and the substantive exercise of independent judgment, exposing a weakness in superficial compliance structures.

Addressing reseacrh question 3, the study identified escalation compression as a major strategic risk. AI-enabled decision support can accelerate threat identification and response planning, but it can also shorten the time available for verification, diplomatic restraint, and command review. This is particularly dangerous when battlefield information is affected by cyber manipulation, spoofed data, deepfakes, or misinformation. Strategic stability in autonomous warfare therefore depends not only on deterrence doctrine, but also on the design of decision support systems that can slow, verify, and audit high-risk decisions.

The main contribution of this study is the Human-Centered Escalation-Aware Governance Framework. The framework integrates eight operational layers human authorization, explainability, auditability, cybersecurity resilience, escalation control, cognitive security, emergency override, and post-action review into a unified decision-support architecture. This contribution advances the intelligent decision support literature by demonstrating that governance must be embedded into the architecture of AI-assisted military decision-making rather than appended as an afterthought, and by translating abstract AI ethics principles into concrete, sequentially activatable control mechanisms.

This study has limitations. It relies on qualitative thematic synthesis conducted by a single researcher and does not include classified military system data, battlefield simulation, expert validation, or empirical testing of autonomous decision support systems. Future research should develop simulation based assessments of escalation-control mechanisms, evaluate human oversight performance under machine-speed decision conditions, test auditability models for AI assisted command systems, and conduct expert validation studies with military practitioners and international lawyers. Further studies should also examine how international organizations and military alliances can operationalize shared standards for meaningful human control across interoperable autonomous systems, and how dynamic layer-activation protocols can be designed to balance governance robustness against operational requirements in diverse threat environments.

Refrences

Amoroso, D., & Tamburrini, G. (2023). Meaningful human control in AI-driven warfare: A legal and ethical framework. *Philosophy & Technology*, *36*(2), Article 25. https://doi.org/10.1007/s13347-023-00678-9

Blanchard, A., Thomas, C., & Taddeo, M. (2024). Ethical governance of artificial intelligence for defence: Normative tradeoffs for principle to practice guidance. *AI & Society*, *40*, 185–198. https://doi.org/10.1007/s00146-024-01866-7

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, *10*, 77110–77122. https://doi.org/10.1109/ACCESS.2022.3191790

Bode, I., Huelss, H., Nadibaidze, A., Qiao-Franco, G., & Watts, T. F. A. (2023). Prospects for the global governance of autonomous weapons: Comparing Chinese, Russian, and US practices. *Ethics and Information Technology*, *25*(1), Article 5. https://doi.org/10.1007/s10676-023-09678-x

Boshuijzen-van Burken, C. (2023). Value sensitive design for autonomous weapon systems: A primer. *Ethics and Information Technology*, *25*, Article 11. https://doi.org/10.1007/s10676-023-09687-w

Boutin, B. (2023). State responsibility in relation to military applications of artificial intelligence. *Leiden Journal of International Law*, *36*(1), 133–150. https://doi.org/10.1017/S0922156522000607

Boyd, J. R. (1996). *The essence of winning and losing*.

Braun, V., & Clarke, V. (2021). One size fits all? What counts as quality practice in (reflexive) thematic analysis? *Qualitative Research in Psychology*, *18*(3), 328–352. https://doi.org/10.1080/14780887.2020.1769238

Christie, E. H., Ertan, A., Adomaitis, L., & Klaus, M. (2023). Regulating lethal autonomous weapon systems: Exploring the challenges of explainability and traceability. *AI and Ethics*, *4*, 229–245. https://doi.org/10.1007/s43681-023-00261-0

Conn, A., & Bode, I. (2025). Establishing human responsibility and accountability at early stages of the lifecycle for AI-based defence systems. *Ethics and Information Technology*, *27*, Article 51. https://doi.org/10.1007/s10676-025-09862-1

Cross, I. C. of the R. (2021). *ICRC position on autonomous weapon systems*. https://www.icrc.org/en/document/autonomous-weapons-icrc-recommends-new-rules

Defense, U. S. D. of. (2023). *DoD Directive 3000.09: Autonomy in weapon systems*. https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Dorton, S. L., & Harper, S. B. (2022). A naturalistic investigation of trust, AI, and intelligence work. *Journal of Cognitive Engineering and Decision Making*, *16*(4), 222–236. https://doi.org/10.1177/15553434221103718

Erskine, T., & Miller, S. E. (2024). AI and the decision to go to war: Future risks and opportunities. *Australian Journal of International Affairs*, *78*(2), 135–147. https://doi.org/10.1080/10357718.2024.2349598

Ferl, A.-K. (2024). Imagining meaningful human control: Autonomous weapons and the (de-)legitimisation of future warfare. *Global Society*, *38*(1), 139–155. https://doi.org/10.1080/13600826.2023.2233004

Floridi, L. (2023). *The ethics of artificial intelligence: Principles, challenges, and opportunities*. Oxford University Press.

Horowitz, M. C., & Kahn, L. (2024). Bending the automation bias curve: A study of human and AI-based decision making in national security contexts. *International Studies Quarterly*, *68*(2), Article sqae020. https://doi.org/10.1093/isq/sqae020

Jenkins, R., Sullins, J. P., Kalu, O., Kamath, A., & Phumjam, K. (2025). Recent insights in responsible AI development and deployment in national defense: A review of literature, 2022–2024. *Journal of Military Ethics*, *24*(1), 63–85. https://doi.org/10.1080/15027570.2025.2483058

Johnson, J. (2021). Artificial intelligence and future warfare: Implications for international security. *Defense & Security Analysis*, *37*(2), 147–169. https://doi.org/10.1080/14751798.2021.1915038

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*(3), 337–359. https://doi.org/10.1017/eis.2021.23

Johnson, J. (2023a). Automating the OODA loop in the age of intelligent machines: Reaffirming the role of humans in command-and-control decision-making in the digital age. *Defence Studies*, *23*(1), 43–67. https://doi.org/10.1080/14702436.2022.2102486

Johnson, J. (2023b). Automating the OODA loop in the age of intelligent machines: Reaffirming the role of humans in command-and-control decision-making in the digital age. *Defence Studies*, *23*(1), 43–67. https://doi.org/10.1080/14702436.2022.2102486

Macrae, C. (2022). Learning from the failure of autonomous and intelligent systems: Accidents, safety and sociotechnical sources of risk. *Risk Analysis*, *42*(9), 1999–2025. https://doi.org/10.1111/risa.13850

Nadibaidze, A., & Miotto, N. (2023). The impact of AI on strategic stability is what states make of it: Comparing US and Russian discourses. *Journal for Peace and Nuclear Disarmament*, *6*(1), 47–67. https://doi.org/10.1080/25751654.2023.2205552

Organization, N. A. T. (2024). *NATO releases revised artificial intelligence strategy*. https://www.nato.int/en/news-and-events/articles/news/2024/07/10/nato-releases-revised-ai-strategy

Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., & others. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, Article n71. https://doi.org/10.1136/bmj.n71

Payne, K. (2021). *I, warbot: The dawn of artificially intelligent conflict*. Oxford University Press.

Rosert, E., & Sauer, F. (2021). How (not) to stop the killer robots: A comparative analysis of humanitarian disarmament campaign strategies. *Contemporary Security Policy*, *42*(1), 4–29. https://doi.org/10.1080/13523260.2020.1771508

Snyder, H. (2023). Systematic literature review as a research method: An updated guide. *Journal of Business Research*, *156*, 113–125. https://doi.org/10.1016/j.jbusres.2022.113125

State, U. S. D. of. (2023). *Political declaration on responsible military use of artificial intelligence and autonomy*. https://www.state.gov/bureau-of-arms-control-deterrence-and-stability/political-declaration-on-responsible-military-use-of-artificial-intelligence-and-autonomy

Timonen, S., & Conlon, C. (2023). Grounded theory in the age of AI: New directions. *Qualitative Research*, *23*(4), 567–584. https://doi.org/10.1177/14687941231168418