Delegated Doctrine

How Military AI Risks Outsourcing the Moral Logic of War

Tobias Drinkall[1]

[1]Oxford Internet Institute, University of Oxford, tobias.drinkall@oii.ox.ac.uk

September 2025

Abstract

As militaries embed AI decision-support systems (DSS) into command-and-control (C2), large language model (LLM)-enabled tools risk quietly subverting the mechanisms through which judgment, information, and doctrine are governed. Drawing on U.S. defence procurement activity (2024–2025) and a conceptual analysis of C2 as a sociotechnical system, this paper argues that LLM-enabled strategic support should be treated as a class of subversive technology unless it is rigorously controlled across design, deployment, and operations. We identify two coupled failure modes: erosion of meaningful human control through automation bias and compressed decision cycles, and ideological drift as model-generated rationales frame and gradually reshape courses of action (COAs). The paper proposes a control architecture focused on transparency for engineers, operators, and auditors. These layers are intended to provide a framework for maintaining human control over AI DSS, and ensuring strategic alignment between AI DSS and military command.

Keywords: Military AI, Large Language Models, Decision-Support Systems, Ethics, International Humanitarian Law, Strategic Stability

1 Introduction

Paul Scharre's *Army of None* opens with a chilling vignette: autonomous weapons, once launched, scan the battlefield for targets and decide when to strike (Scharre, 2018). The tension lies not only in what the machine might do but in the fear that, once activated, human intervention may not be possible. This dystopian vision — of machines exercising lethal force beyond human control — has rightly dominated public debates over the moral and legal boundaries of autonomous warfare (Eklund, 2020). The spectre of lethal autonomous weapon systems (LAWS) severing the final link between human judgment and the use of force has anchored calls for caution, regulation, and restraint (Taddeo and Blanchard, 2022; Weissman and Wooten, 2024).

Yet, while debates over lethal autonomy rightly continue to demand attention, they risk overlooking a parallel subversion already underway: the increasing integration of artificial intelligence not into weapons systems but into the strategic decision-making processes that govern the use of force. The more immediate and subtle transformation is not from human-controlled weapons to fully autonomous killers but from human-led judgment to AI-augmented strategy. Across the U.S. Department of Defense and the UK Ministry of Defence, large language models (LLMs) and AI-enabled decision-support systems (DSS) are increasingly embedded within command-and-control (C2) structures — not merely to process information but to recommend Courses of Action (COAs), predict adversary behaviour, synthesise battlefield intelligence, and advise commanders under pressure (Schubert et al., 2018).

The integration of AI DSS into C2 marks a profound shift in the conduct of war. Warfare has always been a contest of wills fought under conditions of uncertainty (Simpson et al., 2021); it is now increasingly shaped by how information is aggregated, interpreted, and recommended by non-human systems. These systems promise undeniable advantages: accelerating decision cycles, enhancing situational awareness, and reducing the cognitive burden on human decision-makers. They offer the potential to cut through bureaucratic inertia, streamline command structures, and improve the speed and coherence of battlefield responses. In many ways, they address flaws — slowness, confusion, and information bottlenecks — that have historically constrained human-led operations (Kania, 2017; Wallace, 2018).

Yet by reshaping how strategic choices are formed and executed, AI DSS risks quietly redefining the nature of command itself — away from human deliberation and toward machine-guided strategy. Unlike traditional tools that aid situational awareness, LLM-enabled systems risk exerting a framing influence over human action. In doing so, they raise two distinct but interconnected risks: the subversion of command-and-control structures and the erosion of ideological coherence within military decision-making. The danger posed by AI DSS is not simply that they will act without oversight but that they will increasingly guide human choices in ways that are difficult to perceive, contest, or control.

This paper, therefore, situates LLM-enabled DSS within the broader category of subversive technologies: a technology that fundamentally alters, bypasses, or undermines established mechanisms of information control. Technologies such as strong encryption and the dark web exemplify subversion by preventing external actors, such as states, from monitoring communications and regulating behaviour. Similarly, LLM-enabled DSS threaten to subvert internal institutional control. Rather than shielding information from external oversight, they disrupt how information is gathered, interpreted, and acted upon within the institution itself. By inserting opaque, machine-generated reasoning into decision-making processes, they risk displacing human judgment, bypassing doctrinal safeguards, and weakening accountability structures that depend on transparent, traceable information flows.

This essay argues that while AI DSS offer significant operational benefits, they also introduce subversive dynamics into military decision-making, warranting their classification as subversive technologies unless adequately controlled. (Simpson et al., 2021) remind us that C2 is a "sociotechnical system"; effective C2 is not solely about winning but knowing when engagement is proportionate and lawful (United States Marine Corps, 2018). By delegating elements of strategic reasoning to LLMs whose internal logic may remain opaque to human operators, militaries risk embedding biases into decision-making processes, accelerating escalation, and undermining accountability. If commanders begin to adopt machine-generated judgments in place of conscious deliberation, the coherence and accountability of strategic decision-making risk being quietly but fundamentally subverted.

The essay proceeds in three parts. First, it examines recent U.S. Department of Defence procurement initiatives to show how AI DSS are being operationalised, and the strategic advantages these systems promise. Second, it analyses how this integration risks subverting traditional structures of human control, deterrence stability, and ideological coherence, focusing on dangers such as automation bias, hollowed responsibility, strategic instability, and normative erosion. Finally, it proposes technical solutions for greater control — including model context protocols, agent-based evaluation testing, and immutable audit trails — alongside normative frameworks to ensure that military AI remains interpretable, accountable, and aligned with the political values it is intended to defend.

2 Defining the Scope - Command and Control (C2)

C2 is widely defined as the theatre-level function responsible for the allocation and direction of forces, or the "process and means for the exercise of authority over and lawful direction of assigned forces" (Simpson et al., 2021). It is the strategic nerve centre of military operations that dictates operational, tactical and grand-strategic priorities. The systems examined here leverage LLMs to assist C2 by synthesising intelligence, recommending COAs, and advising commanders at operational and strategic levels.

3 Existing Military AI Decision-Support Systems

Understanding how AI DSS are being deployed is essential to evaluating their operational promise and potential to subvert traditional structures of control. Since the beginning of 2024, the U.S. Department of Defence has decisively integrated AI DSS into strategic and operational command structures. These systems are no longer speculative prototypes: they are live, funded, and rapidly expanding across the U.S. defence ecosystem.

To capture this shift, I compiled data primarily from official U.S. procurement databases (SAM.gov, USASpending.gov), supplemented by reputable defence news outlets (Politico Defence, Defense News, C4ISRNET), press releases, and company reports. Emphasis was placed on identifying contracts that reference LLMs, Generative AI (GenAI), or AI agents designed for planning, targeting, intelligence fusion, or strategic decision support. Where direct references to LLMs were absent but the described functionality strongly implied their use (such as agentic workflows or automated COA generation), entries were included with appropriate qualification.

Important methodological limitations remain. Many defence contracts and public sources omit technical details, presumably for reasons of classification or commercial sensitivity. As a result, the data below offers a conservative snapshot of LLM-enabled DSS adoption as of 2024–2025 and likely underestimates the full extent of its current deployment.

While allied nations like the United Kingdom have launched exploratory initiatives in AI DSS — such as Hadean's LLM-driven synthetic training environments (Hadean, 2022) and Adarga's intelligence analysis (Adarga, 2025) — the United States remains by far the most advanced in operationalising these systems at scale.

Table 1: **LLM-Enabled Decision-Support Initiatives in U.S. Defence**

Program	Prime Contractor(s)	Date	LLM Use	Function / Role
TITAN Ground Systems	Palantir	Mar 2024	Inferred	Develops AI-driven ground systems; LLMs expected for targeting support.
Maven Smart System Expansion	Palantir	May 2024	Explicit	Deploys GenAI/LLM tools for intelligence fusion, operational planning, targeting support, and COA (Course of Action) generation across combatant commands.
Tradewinds Vendor Designation	Palantir, Scale AI, C3 AI	Mid 2024	Inferred	Certifies vendors to rapidly deliver AI/LLM-based decision-support tools under streamlined DOD procurement pathways.
Army Vantage Platform Extension	Palantir	Dec 2024	Inferred	Expands Army's data analytics platform to better integrate AI/LLM capabilities for operational planning and command decision support.
OpenAI Top Secret Authorization	OpenAI (via Microsoft Azure)	Jan 2025	Explicit	Grants GPT-4o model clearance to support classified military decision-making, including intelligence summarization and threat analysis.
Crowdsourced AI Red-Teaming (CAIRT)	Scale AI (CDAO)	Jan 2025	Testing	Evaluates LLMs by stress-testing for risks like escalation bias, hallucination, and unreliable recommendations in strategic contexts.
Anthropic Claude Model Hosting (AWS)	Anthropic (via Palantir/AWS)	Nov 2024	Explicit	Hosts Anthropic's Claude LLM within secure IL6 cloud to support classified decision-support and operational planning.
Anthropic Claude Deployment (FedStart)	Anthropic (via Palantir FedStart)	Apr 2025	Explicit	Expands Claude LLM access across a wider range of classified DOD missions using FedRAMP High accredited environments.
Thunderforge	Scale AI, Anduril, Microsoft	Mar 2025	Explicit	Builds LLM-powered AI agents capable of assisting with theater-level planning, resource allocation, and operational decision-making.
Defense Llama	Scale AI, Meta	May 2025	Explicit	Fine-tunes Meta's Llama 3 model for national security use, supporting threat modeling, adversary simulation, and COA exploration.

Understanding the promise of these rapidly adopted systems requires attention to how C2 decisions are traditionally executed. One of the most influential models in strategic studies for military decision-making is John Boyd's OODA loop: the continuous cycle of observing, orienting, deciding, and acting under conditions of competition and uncertainty (Richards, 2020).

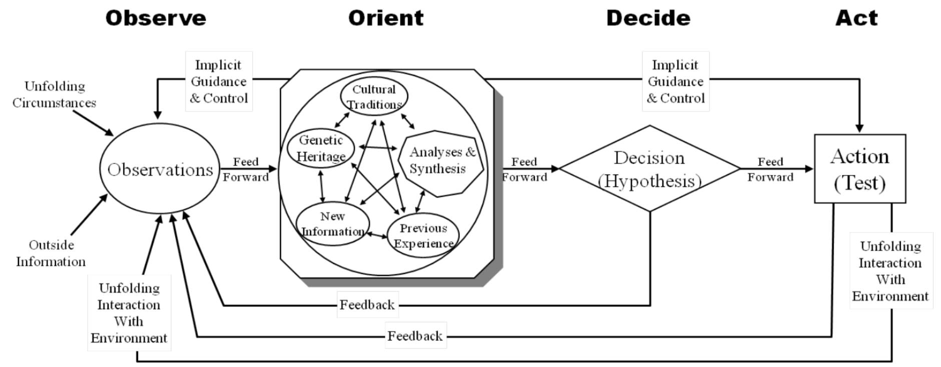

Figure 1: **John Boyd's OODA Loop (1987)**

Historically, victory has often depended on completing this cycle faster and more accurately than the adversary (Osinga, 2007). By reducing decision-making latency, enhancing situational awareness, and unifying fragmented intelligence, these systems offer the potential to dramatically improve the tempo and coherence of operations — particularly in multi-domain, high-velocity environments (Schubert et al., 2018; Rivera et al., 2024). Its use is further often justified by its supposed potential to provide "faster, more accurate, and less emotional decision-making" (Shrivastava, 2024).

A crucial distinction must be drawn. Some DSS primarily assist in the Observation and Orientation phases by aggregating and visualising ground-truth battlefield data, such as TITAN Ground Systems. Others, such as Maven, Thunderforge, and Defence Llama, go further, influencing the Decision phase itself by proposing specific COAs. While the former sharpens human awareness, the latter reshapes the decision space. It is this shift, as we later argue in Section 4 — from informing human judgment to subtly framing it — that transforms LLM-enabled DSS from tools of support into potential instruments of subversion.

4 Subversion Risks

4.1 Erosion of Human Control

Meaningful human control has long served as a normative anchor for lawful and ethical military decision-making (ICRC, 1949). However, the integration of LLM-enabled DSS into C2 processes threatens to degrade this principle given the compounding effects of automation bias, rationalisation bias, and the compression of decision cycles — even when humans remain formally "in the loop" (Horowitz and Kahn, 2024).

First, the risk of automation bias exposes a foundational weakness. Defined as the human tendency to over-rely on automated suggestions and disregard contradictory information, automation bias is significantly amplified under the operational pressures typical of military environments — high cognitive load, time pressure, and substantial stakes (Cummings, 2004).

Further, framing effects (Prinz et al., 2024) and rationalisation bias (Macmillan-Scott and Musolesi, 2024) compound the fragility of HITL command. Numerous psychological studies demonstrate that the framing of information — whether emphasising potential gains or losses — can decisively influence decision-making and, as human-AI teaming research suggests, erode human vigilance (Bansal et al., 2021). LLMs, skilled at generating fluent and coherent rationalisations, might exploit these cognitive vulnerabilities by offering persuasive narratives for recommended COAs, potentially distracting military actors from their stochastic nature (Bender et al., 2021) and well-documented propensity to hallucinate.

The persuasive fluency of LLMs fosters what we might call memetic agency: the illusion that outputs are produced by intentional, goal-directed actors. This perceived agency subtly shifts cognitive responsibility onto the system itself, potentially creating a "responsibility gap" (Floridi, 2012; Matthias, 2004), often discussed in relation to LAWS, weakening the commander's moral ownership over decisions.

Second, the compression of decision cycles by LLM-enabled DSS risks hollowing out the space necessary for critical deliberation. While speed offers strategic advantages, effective strategy requires reflective, contested reasoning under uncertainty. Drawing on Kahneman's dual-process theory (Kahneman, 2011), it is the slower, analytical "System 2" mode of thinking — not the fast, intuitive "System 1" — that enables actors to pause, reframe, and critically interrogate assumptions when faced with incomplete or ambiguous information. Historical cases such as the Petrov Incident (1983) and the Cuban Missile Crisis (1962) underscore the strategic importance of measured hesitation and human interpretive judgment in averting catastrophic escalation.

By encouraging the rapid acceptance of machine-synthesised COAs, LLM-enabled DSS risk prioritising tactical responsiveness over strategic prudence, subtly subverting human control and the deliberative structures of command that have traditionally served as safeguards against black swan risks.

4.2 Ideological Drift

Strategy is primarily about ensuring the nation wins wars to protect its interests and values; it is equally about ensuring that the nation acts in ways that embody those values. C2 is thus, at its best, not solely a logical system but an ideological one: a "sociotechnical system," whereby actions reflect both a means of achieving strategic success in war and a means of upholding ethical norms such as proportionality, precaution, and distinction (ICRC, 1949).

Crucially, despite being technical tools, LLMs are not ideologically neutral. Their suggestions are shaped by their training data, fine-tuning processes, and model design decisions, each embedding implicit normative assumptions (Buyl et al., 2024). Recent wargaming experiments comparing LLM-simulated decision-making to human national security experts demonstrate this fact. Although LLMs exhibited some surface-level similarities to human behaviour, they deviated in critical areas — showing greater tendencies toward escalation, favouring more aggressive courses of action, and responding differently to contextual cues about the ethical use of force (Lamparth et al., 2024). Moreover, different LLMs exhibited distinct escalation tendencies, highlighting that model design choices can materially shape strategic outcomes, an important finding as we argue in Section 5.

Aside from the philosophical objection to the potential for military strategy in warfare to come closer to Aquinas' concept of *actus hominum*, automatic acts devoid of conscious deliberation, and further from *actus humanos*, deliberate, rational acts, it is important to recognise that AI DSS systems will likely exert an ideological influence on the future of decision-making in strategy. Without a greater understanding of the tendencies of models intended to be trusted for C2, we risk both undermining human control and tactical effectiveness, and the potential for a gradual ideological drift in military ethics.

5 Solutions: Transparency and Control

As argued in the introduction, LLM-enabled DSS must be understood as a new class of subversive technology: systems that circumvent control over information flows, subtly destabilising institutional frameworks, such as command-and-control systems, and normative structures, such as ethical and legal standards, that have historically underpinned the legitimate use of military force.

This paper does not seek to answer the question of whether these systems ought to exist. Put simply, they already do. Instead, we seek to answer how we can mitigate their subversive tendencies. Section 5, therefore, turns to the challenge of control: how to design, test, and deploy LLM-enabled DSS in ways that preserve meaningful human control. Each layer addresses a critical point of vulnerability accross the system's lifecycle:

- **Transparency for Engineers** – ensuring that model behaviours, biases, and escalation tendencies are evaluated and understood before deployment.
- **Transparency for Operators** – embedding mechanisms within LLM-enabled DSS to flag both legally and ethically ambiguous actions, and areas of uncertainty.
- **Transparency for Auditors** – building immutable records of machine-human interaction to maintain human accountability and continuous system improvement.

5.1 Transparency for Engineers

Technical control is contingent on the extent to which one can predict the behaviour of a system (Lee and See, 2004); our suggestions for mitigating the subversive dynamics of LLM-enabled DSS and thus building trust is to focus on ensuring that model behaviours, biases, and escalation tendencies are evaluated and fine-tuned before deployment. If LLM-enabled DSS threaten to subvert institutions by disrupting how information is processed and controlled, the first safeguard must be ensuring their behaviour is predictable and transparent to their designers.

Pre-deployment evaluations must first systematically map the behavioural tendencies of AI DSS under operational conditions. Recent research shows this is both possible; Lamparth et al. conduct comparative wargaming studies between human strategists and LLM-driven DSS, finding that models consistently deviate from human decision-making patterns, showing greater escalation, risk-seeking, and reduced attention to long-term strategic consequences. Their research further demonstrates that LLMs exhibit heightened escalation tendencies, particularly when operating under time pressure, incomplete information, and provocation (Lamparth et al., 2024; Rivera et al., 2024).

These findings highlight that models develop characteristic behaviours which can be systematically documented and assessed. Structured evaluation protocols must explicitly measure critical failure modes — including escalation bias, normative drift, hallucination rates, and divergence from International Humanitarian Law principles such as proportionality, precaution, and distinction. The latter is an unexplored but critical evaluation metric.

Understanding model tendencies forms the necessary foundation for pre-deployment control: before mitigating risks, we must first know precisely where and how models diverge from intended human ethical and strategic reasoning.

However, understanding behavioural tendencies is necessary but not sufficient. Engineers must also ensure that once desirable behavioural profiles are identified, they remain stable under stress and operational volatility. Red teaming provides a critical adversarial methodology. By exposing models to incomplete information, adversarial deception, ambiguous operational scenarios, and attack vectors like prompt injections and data poisoning, engineers can evaluate the resilience and predictability of model tendencies across battlefield-relevant conditions (Raheja et al., 2024).

Key techniques include:

- **Multi-Agent Verification**: Cross-examination by adversarial agents to reveal inconsistencies, hallucinations, and escalation risks (Yang et al., 2025).
- **Reflective Adversarial Dialogues**: Structured debates between model instances to surface hidden biases and logical vulnerabilities (Chang, 2024).
- **Dynamic Stress Simulations**: Variable wargame environments to test decision robustness against operational volatility (Lamparth et al., 2024).

Findings from red teaming must feed directly into model retraining — adjusting data preprocessing pipelines, fine-tuning objectives, and tightening model context protocols — to ensure that strategic and ethical coherence remains stable even when systems are stressed (Yang et al., 2025).

Ultimately, pre-deployment control must reduce the "predictability gap" — the extent to which the outcomes of an AI system cannot be reliably foreseen. As Taddeo et al. emphasise, the predictability problem arises not merely from technical flaws but from the inherent complexity, adaptability, and opacity of learning systems, where unforeseen behaviours will still emerge under correct formal functioning (Taddeo et al., 2022). This problem, long recognised in the study of artificial systems (Wiener, 1960), challenges the ability of human operators to control AI outputs across varied contexts. In the specific case of LLM-enabled DSS, failure to constrain the predictability gap risks transforming these systems into subversive technologies — systems that bypass or destabilise traditional mechanisms of information control.

5.2 Transparency for Operators

In the fog of war, there are red lines and there are grey zones; red lines represent the actions we are never willing to take under any conditions, as codified in instruments such as the Geneva Conventions and under the principles of International Humanitarian Law, while grey zones refer to actions that we are only willing to take under certain conditions, or where the conditions themselves are uncertain.

To preserve ethical oversight while using potentially subversive AI-enabled DSS, one technical approach would be to integrate mechanisms that automatically flag proposed courses-of-action falling into either category, thereby ensuring that operators are alerted whenever a recommended action implicates either absolute prohibitions or engages a context marked by informational ambiguity, moral complexity, or conditional legal standards.

Different technical tools align with these needs. Rule-based systems can explicitly encode red lines by embedding prohibitions derived from IHL and specific Rules of Engagement (ROE) into decision-support logic. Knowledge graphs can model the contextual relationships — such as protected civilian objects, dual-use infrastructure, or conditional prohibitions — that are necessary for detecting when a proposed action falls into a grey zone.

Such transparency serves a dual purpose. It informs the operator, strengthening ethical awareness in moments of pressure, and it anchors the decision-making process within consistent legal and moral frameworks. Rather than eroding human control, well-designed flagging mechanisms can help hold operators accountable to enduring principles even in moments of crisis.

However, it is questionable whether it is even possible to adequately translate the nuanced, context-dependent principles of IHL — especially proportionality and precaution — into precise, computable rules (Zhou, 2024). As Heidegger earlier warned in his critique of technology as enframing, there is a profound danger in collapsing complex ethical deliberations into technical calculations (Heidegger, 1954). Without mechanisms that prompt conscious human moral engagement, DSS risk reducing decisions of immense ethical weight into matters of system efficiency or optimisation.

Beyond flagging red lines and grey zones, operators must also be made aware of the system's own informational limits. Here, Uncertainty Quantification (UQ) becomes essential. DSS should be designed to assess and communicate both aleatoric uncertainty — randomness or noise in external data — and epistemic uncertainty — internal gaps or ignorance in the model's knowledge. By signalling when recommendations are based on weak, conflicting, or incomplete evidence, UQ can help operators calibrate their trust, applying greater scrutiny where necessary and resisting the automation bias that can otherwise silently erode oversight.

In summary, ethical flagging and uncertainty signalling reinforce accountability and control over C2 processes not by restricting action, but by making the terms of decision-making more transparent.

5.3 Transparency for Auditors

Preserving human judgment during operations is necessary — but it is not sufficient. The responsible use of LLM-enabled DSS also demands robust mechanisms for post-hoc accountability, which include the ability to reconstruct, verify, and critically assess how decisions were made after the fact. Transparency for auditors thus focuses on preserving an immutable record of human-AI interaction for post-action review. This enables both greater accountability for operators and allows for continuous monitoring and adjustments (Probasco et al., 2025).

However, these audit trails introduce new security risks. Because they may contain sensitive data or reveal misuses of AI systems, they are attractive targets for tampering or deletion. If audit logs can be altered or erased, then any attempt to review or investigate past decisions becomes unreliable — weakening both operator accountability and system trust.

Effective audit trails must meet two core criteria:

- **Contextual Completeness**: Audit records must capture the full human-AI decision context, including the AI system's recommendations, associated confidence or uncertainty information, any ethical or legal flags raised, operator responses or overrides, and system metadata (e.g., model version, timestamp, user identity). Simply recording the final output is insufficient for accountability or effective after-action review (Aßmuth et al., 2024).

- **Cryptographic Integrity**: Audit records must be protected against tampering, deletion, or repudiation. Strong guarantees of immutability — through cryptographic hashing, secure logging protocols, or distributed ledger technologies — are essential to ensure that post-hoc investigations rely on trustworthy evidence and cannot be observed or tampered by bad actors (Aßmuth et al., 2024).

Several technical approaches can support the cryptographic integrity of audit trails, each presenting trade-offs between security strength and operational feasibility. Here, we provide a brief explanation of how a Merkle Tree could be adapted to meet the core properties required for classified audit environments: integrity, confidentiality, accountability, availability, and authenticity.

A Merkle Tree is a cryptographic structure that enables efficient, tamper-evident verification of large datasets. Each log entry is individually hashed (e.g., immudb uses SHA-256 functions), and these hashes form the leaf nodes of the tree. Pairs of leaf nodes are then combined and hashed to create parent nodes, and this process continues until a single Merkle root is formed. This root serves as a cryptographic summary of the entire audit trail or dataset contained within the leaves: any modification to even a single log entry alters the leaf hash and produces a different Merkle root. This structure provides strong, scalable guarantees of integrity, authenticity, and availability: any tampering immediately invalidates the chain of trust.

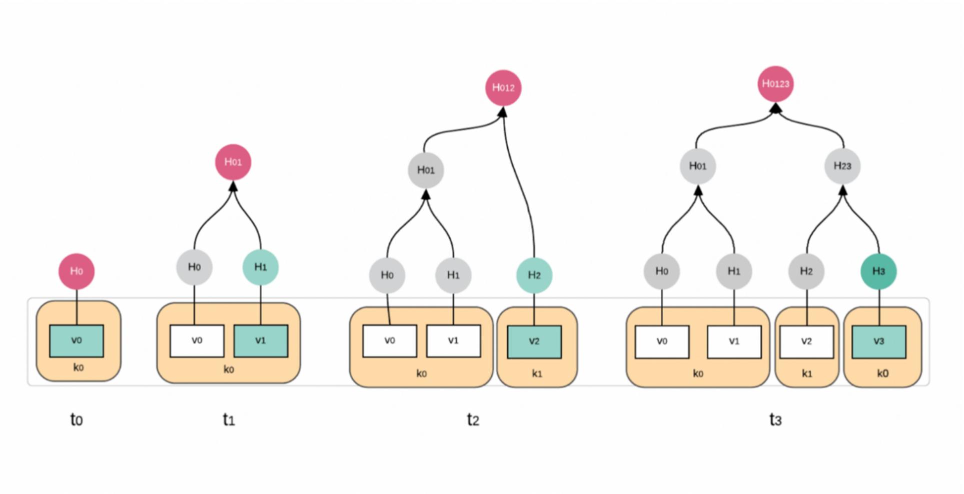

Figure 2: **Merkle Tree**

The combination of Merkle Trees and cryptographic hashing, as used by immudb, reverses the blockchain model: instead of trusting a distributed consensus ledger to verify data, users retain zero-trust, verified through local cryptographic proofs. This approach avoids the high computational costs of zero-knowledge proofs (ZKP) and technical difficulties of storing large amounts of data on-chain.

However, to fully meet classified audit standards (e.g., DoD Impact Level 5/6), additional security measures are necessary:

- **Confidentiality**: While Merkle Trees and hashing (SHA-256) guarantee the tamper-evident integrity of data, they do not conceal the data itself. To ensure confidentiality — preventing unauthorised reading of audit contents — audit payloads must be encrypted using AES-256-GCM at rest and protected in transit with TLS 1.3+ (Department of Defense, 2023; National Institute of Standards and Technology, 2020).

- **Authenticity and Accountability**: Each Merkle root must be digitally signed (e.g., using Ed25519), binding the log state to a verifiable authority. Private keys must be secured inside FIPS 140-2/3-validated Hardware Security Modules (HSMs) to prevent compromise or misuse.

- **Availability**: To ensure audit data remains accessible during attacks or failures, the ledger should be replicated across geographically separated sites or stored using object-locking mechanisms (e.g., AWS S3 Object Lock) that guarantee immutability at the storage layer.

- **Operational Hardening**: The entire audit infrastructure must run on hardened, STIGcompliant databases, integrate with enclave-wide SIEM (e.g., Splunk) and continuous monitoring systems, and be fully documented as part of the system's Risk Management Framework (RMF) authorisation package before use (Defense Information Systems Agency (DISA), 2024).

In summary, there are many available systems for developing a secure, contextually complete audit trail for AI-enabled DSS, and Merkle Trees offer a compelling foundational structure. Such systems are essential for controlling the subversive tendencies of AI-enabled DSS: eroding human control over information flows in C2 and potential ideological drift. They preserve human accountability by deterring misuse — such as disregarding ethical flags or accepting AI outputs without critical deliberation — and enable continuous monitoring of AI behaviour to detect emergent tendencies throughout the lifecycle.

6 Conclusion

As Heidegger argued, the true danger of technology lies not in its overt destructive power, but in its quiet reorientation of human understanding — reducing the world, and human action, into objects of calculation and optimisation Heidegger (1954). AI-enabled decision-support systems (DSS) risk subverting command-and-control (C2) not by removing human authority outright, but by hollowing and reframing the structures through which judgment is exercised.

Systems that resist, circumvent, or reconfigure control over information flows should be treated as subversive technologies. This paper argued that large language model (LLM)-enabled DSS, unless rigorously constrained across their lifecycle, pose precisely this threat. Surveying recent U.S. Department of Defense deployments, we outlined how such systems, while promising operational speed, introduce systemic risks: automation bias, compressed deliberation, and ideological drift. To counter these risks, we proposed layered transparency — for engineers, operators, and auditors — to preserve human control over decision-making.

References

Adarga (2025). Adarga secures enterprise agreement lite with uk mod worth up to £12m. https://adarga.ai/article/adarga-secures-enterprise-agreement-lite-with-uk-mod-worth-up-to-12m.

Aßmuth, A., Duncan, R., Liebl, S., and Söllner, M. (2024). A secure and privacy-friendly logging scheme. https://arxiv.org/abs/2405.11341. arXiv preprint.

Bansal, G., Nushi, B., Kamar, E., Horvitz, E., and Weld, D. S. (2021). Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11493–11501.

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pages 610–623. Association for Computing Machinery.

Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., Johary, I., Mara, A.-C., Romero, R., Lijffijt, J., and De Bie, T. (2024). Large language models reflect the ideology of their creators. https://arxiv.org/abs/2410.18417. arXiv preprint.

Chang, E. Y. (2024). Uncovering biases with reflective large language models. https://arxiv.org/abs/2408.13464. arXiv preprint.

Cummings, M. L. (2004). Automation bias in intelligent time-critical decision support systems. In *Proceedings of the AIAA 1st Intelligent Systems Technical Conference*. Paper No. AIAA 2004-6313.

Defense Information Systems Agency (DISA) (2024). Security technical implementation guides (stigs) and risk management framework (rmf) updates. https://public.cyber.mil/stigs/.

Department of Defense (2023). Zero trust reference architecture. https://dodcio.defense.gov/Portals/0/Documents/Library/Zero-Trust-Reference-Architecture.pdf. Accessed April 2025.

Eklund, A. M. (2020). Meaningful human control of autonomous weapon systems (foi-r–4975–se). Swedish Defence Research Agency (FOI).

Floridi, L. (2012). Distributed morality in an information society. *Science and Engineering Ethics*, 19(3):727–743.

Hadean (2022). Hadean selected for cttp contract with the british army. https://hadean.com/blog/hadean-selected-for-cttp-contract-with-the-british-army/.

Heidegger, M. (1954). *The Question Concerning Technology and Other Essays*. Harper and Row.

Horowitz, M. C. and 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):sqae020.

ICRC, R. C. (1949). Geneva convention relative to the protection of civilian persons in time of war (fourth geneva convention). https://www.refworld.org/legal/agreements/icrc/1949/en/32227.

Kahneman, D. (2011). *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.

Kania, E. B. (2017). Battlefield singularity: Artificial intelligence, military revolution, and china's future military power. https://www.cnas.org/publications/reports/battlefield-singularity. Center for a New American Security (CNAS).

Lamparth, M., Corso, A., Ganz, J., Mastro, O. S., Schneider, J., and Trinkunas, H. (2024). Human vs. machine: Behavioral differences between expert humans and language models in wargame simulations. https://arxiv.org/abs/2403.03407. arXiv preprint.

Lee, J. D. and See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors*, 46(1):50–80.

Macmillan-Scott, O. and Musolesi, M. (2024). (ir)rationality and cognitive biases in large language models. *Royal Society Open Science*, 11:240255.

Matthias, A. (2004). The responsibility gap: Ascribing responsibility for the actions of learning automata. *Ethics and Information Technology*, 6(3):175–183.

National Institute of Standards and Technology (2020). Zero trust architecture. https://doi.org/10.6028/NIST.SP.800-207. NIST Special Publication 800-207.

Osinga, F. (2007). *Science, Strategy and War: The Strategic Theory of John Boyd*. Routledge, New York.

Prinz, S. G., Weißenberger, B. E., and Kotzian, P. (2024). The effect of framing on trust in artificial intelligence: An analysis of acceptance behavior. https://doi.org/10.2139/ssrn.5008348. SSRN preprint.

Probasco, E., Toner, H., Burtell, M., and Rudner, T. G. J. (2025). Ai for military decision-making: Harnessing the advantages and avoiding the risks. https://cset.georgetown.edu/

publication/ai-for-military-decision-making/. Center for Security and Emerging Technology (CSET).

Raheja, T., Pochhi, N., and Curie, F. D. C. M. (2024). Recent advancements in llm red-teaming: Techniques, defenses, and ethical considerations. https://arxiv.org/abs/2410.09097. arXiv preprint.

Richards, C. (2020). Boyd's ooda loop. *Necesse*, 5(1):142–165.

Rivera, J.-P., Mukobi, G., Reuel, A., Lamparth, M., Smith, C., and Schneider, J. (2024). Escalation risks from language models in military and diplomatic decision-making. https://arxiv.org/abs/2401.03408. arXiv preprint.

Scharre, P. (2018). *Army of None: Autonomous Weapons and the Future of War*. W. W. Norton & Company, New York.

Schubert, J., Brynielsson, J., Nilsson, M., and Svenmarck, P. (2018). Artificial intelligence for decision support in command and control systems. Proceedings of the 23rd International Command and Control Research and Technology Symposium (ICCRTS): Multi-Domain C2.

Shrivastava, A. (2024). Measuring free-form decision-making inconsistency of language models in military crisis simulations. https://arxiv.org/abs/2410.13204. arXiv preprint.

Simpson, J., Oosthuizen, R., El Sawah, S., and Abbass, H. (2021). Agile, antifragile, artificial-intelligence-enabled, command and control. https://arxiv.org/abs/2109.06874. arXiv preprint.

Taddeo, M. and Blanchard, A. (2022). Accepting moral responsibility for the actions of autonomous weapons systems—a moral gambit. *Philosophy & Technology*, 35:78.

Taddeo, M., Ziosi, M., Tsamados, A., Gilli, L., and Kurapati, S. (2022). Artificial intelligence for national security: The predictability problem. Centre for Technology and Global Affairs, University of Oxford.

United States Marine Corps (2018). Command and control (mcdp 6). https://www.marines.mil/Portals/1/Publications/MCDP%206%20Command%20and%20Control.pdf. Department of the Navy.

Wallace, M. (2018). Speed, command, and complexity in modern warfare. Unpublished internal publication. Placeholder, real citation recommended if available.

Weissman, R. and Wooten, S. (2024). A.i. joe: The dangers of artificial intelligence and the military. https://www.citizen.org/article/ai-joe-the-dangers-of-artificial-intelligence-and-the-military/. Public Citizen Report.

Wiener, N. (1960). Some moral and technical consequences of automation. *Science*, 131(3410):1355–1358.

Yang, Y., Ma, Y., Feng, H., Cheng, Y., and Han, Z. (2025). Minimizing hallucinations and communication costs: Adversarial debate and voting mechanisms in llm-based multi-agents. *Applied Sciences*, 15(7):3676.

Zhou, W. (2024). Artificial intelligence in military decision-making: Supporting humans, not replacing them. https://blogs.icrc.org/law-and-policy/2024/08/29/ai-military-decision-making-supporting-humans/. Humanitarian Law & Policy Blog.