Regulating LLMs in Warfare

Naomi Solomon
Stanford University

Executive Summary

Large language models (LLMs) are moving from general-purpose tools into defense workflows that support intelligence analysis, planning, logistics, cyber operations, and information activities. Yet the United States lacks clear, enforceable standards that specify when LLM outputs may be used in high-consequence military contexts, what testing and monitoring are required before deployment, and how accountability attaches when LLM-assisted decisions contribute to harm or escalation. LLMs currently exist in a governance gap that can (1) amplify escalation risk through overconfident or brittle recommendations under uncertainty, (2) lower the cost and scale of information operations and disinformation, and (3) expand the attack surface for adversarial manipulation, including data poisoning and sensitive-data leakage.

This memo proposes a U.S. regulatory framework for military LLM accountability organized around four pillars:

- **Human decision rights and escalation controls**: Require documented human authorization for specified operational decisions and establish escalation monitoring for any LLM used in strategic or crisis-sensitive contexts.

- **Human approval for information operations content**: Require mandatory human review and approval for AI-generated content intended for information warfare or influence activities.

- **Security, data protection, and adversarial testing**: Set baseline requirements for secure training and deployment pipelines, controlled data handling, and continuous red-teaming against model- and data-centric attacks.

- **Accountability and traceability mechanisms**: Mandate audit trails for LLM outputs used in operational contexts, clarify responsibility assignments to human decision-makers, and establish formal incident reporting and review procedures.

- **In-scope uses**: intel synthesis, planning support, logistics, cyber support, IO content drafting, translation, summarization, and related functions.

By translating responsible AI principles into operationally enforceable safeguards, the U.S. can reduce avoidable escalation and security failures, strengthen civilian oversight of military AI, and shape credible norms for responsible use of LLMs in defense settings.

I. Introduction

As large language models (LLMs) grow more advanced with each new update, their potential for use in military contexts is expanding at a rapid rate. These models have the ability to generate human-like text, analyze intelligence data, and assist military officials with complex decision-making tasks. Integration into military operations is no longer hypothetical, but already under testing. In 2024–2025, DoD and the Army moved from exploratory discussion to structured generative-AI adoption, including the CDAO's AI Rapid Capabilities Cell funding GenAI-focused pilots and user-centered experimentation, and the Army's launch of an enterprise LLM workspace alongside public work on AI-enabled support for command-and-control.

However, despite the fact that these models are being deployed in high-stakes environments, these systems continue to operate in a regulatory vacuum. Unlike autonomous drones or conventional weapons systems, LLMs are still not governed by any explicit military protocols. This absence of clear oversight introduces a number of serious risks, such as the misinterpretation of sensitive geopolitical situations, the escalation of conflict through AI-generated decisions, and the accountability gaps that can arise when decisions have unintended consequences.

A recent, widely reported case illustrates the governance risks that arise when AI-enabled systems are used in high-consequence military workflows without clear, enforceable oversight. In early 2024, reporting described the Israeli military's use of an AI-assisted targeting system ("Lavender") in Gaza and raised concerns about the speed of target nomination, the adequacy of human review, and the potential for erroneous identification under operational pressure. The ensuing scrutiny from civil society and international officials underscored a broader point: when AI tools are integrated into sensitive decision processes without transparent standards for authorization, auditing, and accountability, the probability of operational error and strategic blowback increases.

Large language models differ from many other military AI systems because they are general-purpose, easy to redeploy across missions, and capable of producing persuasive outputs at scale for planning, intelligence synthesis, cyber, and information operations. Those properties make LLMs especially susceptible to misuse, overreliance, and adversarial manipulation unless

they are bounded by explicit decision rights, testing requirements, and traceable approval procedures.

The policy proposed in this paper aims to close this current regulatory gap by establishing specific safeguards for the use of LLMs in defense contexts. These regulations include mandated human oversight, content approval from humans in information warfare, and defined accountability systems. In such a protocol, the U.S. would enhance the safety and reliability of its own operations and decisions while also setting a global precedent for ethical military AI governance around the globe, pushing other nations to model their frameworks similarly to promote international stability.

II. Key Risks of LLMs in Warfare

Unintentional Escalation in LLM Decision-Making

To standardize how these risks are identified, measured, and managed, this memo aligns with NIST's AI Risk Management Framework (AI RMF 1.0), which organizes AI risk work across governance, context-mapping, measurement, and ongoing risk management. For generative systems such as LLMs, NIST's Generative AI Profile further specifies risk categories and evaluation considerations that are novel to or exacerbated by generative AI. Together, these references provide a common vocabulary for the risk categories below and support translating them into auditable controls later in the memo.

Strategic recommendations that are generated by LLMs can escalate conflicts in ways that humans deploying the technology did not intend. A policy brief by Stanford's Human-Centered AI Institute found that every major LLM tested in wargame simulations displayed similar patterns of unpredictable escalation. In some cases, the models went so far as to recommend nuclear strikes based on misinterpretations of adversary behavior.

These systems are not yet able to reliably interpret the complex and nuanced signals that are inherent in modern geopolitical conflicts. Misreading subtle signals such as intent, tone, or context could trigger a chain of military actions that the human in charge did not intend. This error rate is a problem even when these models are explicitly trained and fine-tuned for military use. Performance benchmarks are often reached in ideal conditions that fail to reflect the unpredictability of real-world conflict zones.

Human authorization gate (crisis-sensitive/strategic contexts). For any LLM used in crisis-sensitive or strategic planning contexts, a human authorizing official (O-5+ or designated civilian equivalent) must approve outputs before operational use; approval must be recorded in a decision log containing the prompt, the output, and a brief justification.

Lower Barrier to Entry into Information Warfare

LLMs are deployed widely by both state and non-state actors. Tools like ChatGPT and Llama, as well as smaller open-source models, easily generate persuasive and widespread propaganda, misinformation, or psychological operations. This vulnerability makes it significantly easier for non-state actors to wage information warfare, lowering the barrier to entry for adversarial actors who previously would have lacked the resources or the expertise to effectively engage in these activities.

Such accessibility also tempts state actors who are authorized to conduct psychological operations to scale up campaigns, without proportionally increasing oversight. A single operator using an LLM can now generate thousands of messages a day, which is significantly more efficient than the traditional methods, and this raises the risk that overly harmful or inflammatory content slips through human review.

To mitigate these concerns, all AI-generated content intended for information warfare must go through mandatory human review and approval before deployment.

Increased Vulnerability to Security Breaches

Using LLMs in military workflows expands the attack surface across training pipelines, model supply chains, retrieval corpora, and deployed interfaces. Models can be compromised through data poisoning (training or retrieval), prompt-injection and tool-exfiltration attacks, or supply-chain manipulation of weights and dependencies. One of the most concerning risks is dataset poisoning, where malicious actors subtly manipulate a model's training data in order to alter its behavior in undetectable ways, and could lead to decisions or outputs that might appear reasonable, but are compromised.

Another major threat is model inversion, which is a form of attack where adversaries extract sensitive training data by analyzing the model's outputs. For LLMs trained on classified or confidential military documents, this vulnerability could result in serious security breaches, without any visible indicators that the LLM has been compromised.
The military needs to adopt strict and uniform data protection standards, requiring secure and vetted training pipelines and implementing continuous adversarial testing before field deployment.

Lack of Accountability Structures

The use of LLMs in decision-making processes complicates the question of responsibility. When an AI-generated decision, or even recommendation, leads to unintended harm, it becomes unclear who is responsible for bearing the legal and ethical consequences. Is it the developer, the user, the commander? Lack of clarity about who is accountable for mistakes, undermines trust in

these systems and also incentivizes riskier decision-making. If operators feel shielded by the fact that an AI is the one that made the harmful decision, military actors might act recklessly. Without clear accountability structures, officials can attribute unethical outcomes to the mysterious, black-box system.
To avoid the breakdown in responsibility, the military must enforce traceability for all AI-enabled decisions, establishing protocols that clearly assign accountability to human actors.

These risks that are posed by unregulated LLMs are growing with every AI advancement. Without urgent regulatory action, these systems have the power to undermine trust in these operational systems, lower the threshold for conflict, and erode public trust in military decision-making

III. Current Approaches

The Department of Defense's primary binding policy on autonomy in weapon systems, DoD Directive 3000.09, establishes requirements for human judgment, senior-level review, and risk mitigation in the development and use of autonomous systems. However, this directive was developed primarily to govern autonomous weapon systems and does not fully address the growing use of large language models as decision-support tools that shape targeting, escalation, and operational judgment upstream of lethal action.

Both international and national entities have started to build frameworks for AI in military contexts, but none have meaningfully accounted for the unique risks posed by large language models used in warfare. This section evaluates two of the most prominent efforts, the UN's Convention on Certain Conventional Weapons and the U.S. Department of Defense's AI principles, and identifies the critical gaps that they leave open.

United Nations: Convention on Certain Conventional Weapons (CCW)

The most relevant international effort to address autonomous military systems has come through the UN's Convention on Certain Conventional Weapons (CCW), specifically through its Group of Governmental Experts (GGE) on Lethal Autonomous Weapons Systems (LAWS). These experts have been tasked with developing a framework that addresses the emerging technologies in warfare, including AI-driven systems. The 2024 GGE sessions focused on gathering state input, reviewing legal and ethical standards, and drafting a potential framework that could guide international governance of lethal autonomous systems.

However, this process remains slow-moving, non-binding, and limited in scope. The GGE's work primarily targets lethal autonomous weapons, not large language models. LLMs, which tend to be non-lethal, and are deployed in intelligence, communication, or cyber operations, fall outside the limited operational focus of lethal autonomous weapons. Because of this, the CCW

framework offers little clarity or enforceable standards on the regulation of LLMs in warfare, leaving their numerous risks unaddressed in global governance efforts.

United States: Department of Defense AI Principles

On the domestic front, the U.S. Department of Defense has adopted a set of ethical principles for artificial intelligence, developed by the Defense Innovation Board and issued in 2020. These principles emphasize the need for responsible, equitable, traceable, reliable, and governable AI systems. They aim to ensure that AI technologies are tested, auditable, and compliant with the laws of war, before they are deployed. The Department also established what is now known as the Chief Digital and Artificial Intelligence Office (CDAO) to coordinate the implementation of these standards across different branches.
These values provide a useful foundation, but they are not operationalized into enforceable requirements for evaluation, monitoring, logging, and lifecycle controls in real-world military use. They also do not directly address LLM-specific risks such as hallucinated or miscontextualized intelligence, prompt-injection and tool misuse, training-data compromise, and large-scale information operations. Without requirements tailored to these failure modes, the Principles alone are insufficient to govern LLM deployment in defense settings.

The result is a regulatory gap: international discussions largely remain oriented toward lethal autonomous weapons and consensus-based, non-binding outcomes, while U.S. policy articulates high-level ethical principles without translating them into LLM-specific operational controls. Meanwhile, LLMs are being tested and integrated into planning, intelligence analysis, and information operations in ways that can materially shape targeting and escalation decisions. Without dedicated guardrails, the risk of inappropriate reliance, compromised outputs, and escalation pathways increases as deployment scales.

To close this gap, the United States needs to take the lead in establishing a specific regulatory framework, designed for LLMs in military operations. The next section outlines a targeted set of regulations that respond directly to the risks discussed above and offers a roadmap for ensuring an accountable and ethical deployment of LLMs in national defense.

IV. Proposed Regulation

To address the urgent risks of LLMs in military operations, the United States must lead in establishing a concrete and enforceable regulatory framework. This framework includes specific operational rules, technical standards, and accountability structures, as well as the creation of an oversight body to ensure consistent and uniform compliance. Below are four key areas of regulation that form the foundation of this policy.

Mandated Human Oversight and Escalation Monitoring

No autonomous operational use in high-consequence contexts. An LLM may not independently initiate, approve, or execute (a) the use of force, (b) target nomination/prioritization, (c) changes to rules of engagement, or (d) crisis-sensitive operational plans. In these contexts, LLM outputs may be deployed only as decision-support and require documented approval by a designated human authorizing official before operational use.

Additionally, appropriately cleared assessors should conduct periodic independent audits to assess whether the LLMs are performing within the expected ethical and strategic boundaries. These audits must also include escalation-risk testing using structured simulations and wargame-like scenarios. If an LLM fails to pass these escalation-responsiveness thresholds, its deployment should be immediately suspended.

Approval of AI content in Information Warfare

All AI-generated content that is intended for psychological operations, propaganda, or other forms of information warfare must go through multi-tiered human reviews. Each deployment must be cleared by at least two levels of review: one at the individual operator level and another by a designated review board that contains both AI and information warfare experts.

IO content control, classification, and audit trail. Any LLM-generated content intended for psychological operations, propaganda, influence activities, or deception must be (1) labeled and classified at creation, (2) reviewed and approved through the existing two-level process, and (3) entered into an IO Content Log that stores the prompt, output, intended audience/channel, approving officials, and the operation identifier. Synthetic media (including deepfakes) may be produced only under a written authorization tied to a specific operation identifier and must include embedded provenance/watermarking and distribution controls so the content can be traced and audited.

A potential way for compliance with this policy to be enforced is to integrate digital watermarking and other content tracking systems that have the ability to tag AI-generated outputs, ensuring traceability of the content and preventing unauthorized dissemination.

Strengthening AI Data Protections

All LLMs that are deployed in military settings must be trained on pre-cleared and classified datasets that have undergone rigorous vetting for integrity, origin of the data, and adversarial manipulation. These datasets should be stored in secure, controlled enclaves and encrypted with defense-grade standards that are held uniform across military sectors.

Before deployment, each model must undergo pre-deployment adversarial evaluation including prompt-injection, jailbreak resistance, sensitive-data leakage, tool-misuse scenarios, and poisoning/compromise detection. The findings from the red-teaming and testing must be

submitted to a registry maintained by the oversight committee outlining exactly which tests were conducted.
The risk monitoring should be ongoing, including detection systems that are able to identify unusual shifts in model output patterns that could indicate that there was post-deployment tampering of the model, or covert retraining attempts.

Accountability measures for AI decision-making

An independent Military AI Oversight Committee (MAIOC) should be established under the Department of Defense, with cross-branch representation among a number of interdisciplinary experts, ranging among AI specialists, military legal experts, intelligence officers, and ethics advisors.

The committee will have the authority to enforce compliance with AI regulations outlined in this paper through unplanned inspections and audits, suspensions of AI systems that were planned to be deployed based on these risk assessments, the mandatory removal of a system when the oversight protocols are violated, investigations into LLM-related incidents, and publishing accountability reports.

MAIOC will also maintain a centralized, encrypted log of all LLM-generated military content, decisions, and deployment records. Every AI-generated output must be timestamped, attached to a supervising human operator, and stored for at least five years. Chain-of-command responsibility must be explicitly documented and enforced. If an LLM's recommendation is used, the human approver bears legal and ethical responsibility for that outcome.

This structure ensures that human actors are not able to deflect blame onto an AI system, reinforcing accountable decision-making throughout the entire chain of command.

As a global leader in both military power and technological innovation, the United States is uniquely positioned to set the standard for responsible AI use in warfare. By implementing these specific regulations, the U.S. can help establish a secure and ethical precedent that other nations can adopt to reduce global instability and misuse of emerging military technologies.

V. Stakeholders

Successfully implementing a regulatory framework for LLM use in military operations will require the support of a diverse and broad collection of stakeholders. Ranging from government agencies to academic institutions and advocacy groups, these organizations have overlapping interests and priorities in ensuring that military AI systems are secure, accountable, and ethically governed.

Department of Defense (DoD)

The DoD is necessary to the implementation and enforcement of these regulations. It already has a foundation of ethical AI principles aimed at responsible use, but integrating LLM-specific norms would give it further ethical guidance, reducing any ambiguity in deployment protocols. The DoD also has a vested interest in reducing national security breaches, avoiding strategic mistakes, and preserving U.S. leadership in AI military ethics. These regulations would provide institutional clarity and a clear risk mitigation structure, both of which are necessary for safe and effective LLM adoption across the military branches.

Intelligence Agencies (NSA, CIA)

Agencies such as the National Security Agency (NSA), Central Intelligence Agency (CIA), and Defense Intelligence Agency (DIA) have strong incentives to support this LLM regulation. These institutions rely on large-scale language processing for tasks such as intelligence gathering, detection of disinformation, cybersecurity, and threat analysis. Poorly regulated LLMs present dangerous risks to information safety, operational security, and data exposure. Structured regulation would help these agencies develop LLMs that can improve efficiency in a responsible way, ensuring safeguards against adversarial manipulation, data leaks, and escalation from misinterpretation.

Human-Centered AI Research Institutes

AI safety research institutes are also well-positioned to support and inform military LLM regulation. Research centers and institutes such as Stanford HAI, UC Berkeley's Center for Human-Compatible AI, and Georgetown's CSET would be able to offer technical expertise to help design the benchmarks, audits, and testing environments that are needed to keep these LLMs secure, ensuring that military LLMs are held to high enough safety standards.

Ethics and Human Rights Organizations

Groups such as Human Rights Watch, Amnesty International, AlgorithmWatch, and the Center for AI and Digital Policy (CAIDP) share the mission of monitoring and preventing abuses in the use of emerging technologies. They are all likely to support stronger oversight of military AI systems, especially models that are able to influence information warfare and decision-making. These organizations would be likely to support LLM regulation to protect international humanitarian law and reduce risks to civilian populations in conflict zones.

Tech Policy Think Tanks and Non-Profits

A number of tech policy organizations would be allies in advancing this regulatory framework. Groups like the Center for Democracy and Technology (CDT), the Brookings Institution, the Center for a New American Security (CNAS), and the Partnership on AI all work at the intersection of technology, national security, and ethics. These institutions help shape regulatory policy, analyze potential impacts of tech implementation, and try to build bipartisan support

among lawmakers and defense leadership. Their involvement would lend intellectual legitimacy and political momentum to this proposed policy.

These stakeholders would all be interested in ensuring that the LLMs that are deployed in military contexts are used with caution, oversight, and accountability. Their support would be essential in advancing a regulatory framework that balances innovation with national and global security.

VI. Anticipated Opposition

While there are many stakeholders who would support the regulation of LLMs in military contexts, there are also a number of organizations who may resist certain elements of this proposed policy. Understanding their concerns is essential for developing a strategy that encourages a compromise without diluting the policy's core safety principles.

Defense Technology Contractors

Major defense tech contractors like Palantir, Raytheon Technologies, and Lockheed Martin may view these regulations as an unnecessary obstacle to rapid innovation and deployment of their technologies. These companies are experiencing constant pressure to push out AI-enabled tools faster than their adversaries, and strict oversight regulations could be seen as bureaucratic slowdowns. In particular, mandated audits, human-in-the-loop requirements, and deferring to an oversight committee may be perceived as too strict of requirements.

Additionally, the contractors involved in LLM development may resist any resulting training data transparency rules or mandatory red-teaming, as this regulation could put their proprietary models or trade secrets at risk.

The policy should allow exceptions so that oversight could occur without compromising sensitive intellectual property for the companies. Additionally, involving defense contractors early in the regulatory design process could also encourage buy-in while preserving the need for developer accountability.

Military Leaders

Military leadership may view these LLM regulations as encroachments on their strategic autonomy. Commanders in fast-moving environments may see these mandatory human-in-the-loop checks or logging requirements as sources of friction that stop them from doing their jobs effectively. There may also be a certain level of skepticism toward non-military oversight committees.

Such opposition can be addressed by embedding military leadership within the oversight committee itself, ensuring that the people who are enforcing these safety standards are doing so with an understanding of how they would translate to real military situations. The regulation should also emphasize that the intent of this regulation is simply to ensure that AI use in warfare remains lawful, strategic, and accountable, not to place unnecessary restrictions.

Tech Libertarians and Anti-Regulatory Policymakers

Policymakers who are aligned with tech-libertarian or deregulatory ideologies may find an issue with any new government constraints on innovation, arguing that regulation will simply snuff out U.S. competitiveness in the AI arms race, or open the door to broader constraints on military modernization.

It is important to frame regulation as pro-security, rather than anti-innovation. Rather than banning capabilities, the policy proposal focuses on ensuring that those capabilities don't lead to escalation, misuse, or accidental war. Emphasizing that these standards protect those involved in the military operations, preserve public trust, and align with existing DoD ethics principles can help shift the perspective away from one of anti-innovation.

Some organizations and companies may resist the regulation out of concern that it would slow innovation or limit autonomy. However, without clear rules, there is a heightened risk for security and ethical concerns. Regulation offers legal clarity, operational stability, and a safeguard against escalating liabilities towards both deployers and developers. The goal is to ensure that innovation occurs within a regulatory framework that protects global stability.

Conclusion

LLMs will shape military operations whether or not governance keeps pace. The policy question remains therefore how to deploy large models under a coherent accountability regime that preserves human responsibility, reduces escalation risk, and resists adversarial manipulation. Today's governance tools remain incomplete for LLM-enabled workflows: international efforts largely focus on lethal autonomous weapons and move slowly, while U.S. policy emphasizes broad ethical principles without converting them into specific, enforceable requirements tailored to LLM risks.

This memo establishes minimum safeguards for the most foreseeable failure modes of LLMs in warfare, including escalation through misinterpretation, scalable information operations, and cyber vulnerabilities such as data poisoning or sensitive-data extraction. A workable baseline should (1) draw bright lines around restricted and prohibited uses; (2) require auditable testing, documentation, and lifecycle monitoring; (3) mandate security controls proportionate to

adversarial threat models; and (4) create an oversight mechanism that learns from incidents and updates standards on a predictable cycle.

Critics may argue that lethal autonomous weapons deserve priority. That is correct, and it is not a reason to defer LLM governance. LLMs operate in adjacent, high-impact domains (planning, intelligence, cyber, and information operations) where failures can still produce strategic consequences and accountability gaps. Establishing LLM-specific guardrails now complements existing weapons autonomy policy and strengthens U.S. credibility when advocating for responsible military AI internationally.

References

American Forces Communications and Electronics Association. (2024, August 20). *U.S. Army gives update on generative artificial intelligence and large language models*. https://www.afcea.org/signal-media/us-army-gives-update-generative-artificial-intelligence-and-large-language-models

Business Insider. (2024, April). *Israel using AI in Gaza targets offers terrifying glimpse at future of war*. https://www.businessinsider.com/israel-using-ai-gaza-targets-terrifying-glimpse-at-future-war-2024-4

Department of Defense. (2023, January 25). *DoD Directive 3000.09: Autonomy in weapon systems* (DoDD 3000.09). https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf

Defense Innovation Board. (2020, February 24). *DoD adopts ethical principles for artificial intelligence*. U.S. Department of Defense. https://www.defense.gov/News/Releases/release/article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence/

Financial Times. (2024, April). *Israel's use of AI in Gaza raises concerns about oversight*. https://www.ft.com/content/f6c1b6ab-532c-44b0-9caa-7c8714c85764

Hogan Lovells. (2024). *Model inversion and membership inference: Understanding new AI security risks and mitigating vulnerabilities*. https://www.hoganlovells.com/en/publications/model-inversion-and-membership-inference-understanding-new-ai-security-risks-and-mitigating-vulnerabilities

Olanipekun, S. O. (2025). Computational propaganda and misinformation: AI technologies as tools of media manipulation. *World Journal of Advanced Research and Reviews, 25*(1), 911–923. https://doi.org/10.30574/wjarr.2025.25.1.0131

SentinelOne. (n.d.). *Data poisoning*. Retrieved January 11, 2026, from https://www.sentinelone.com/cybersecurity-101/cybersecurity/data-poisoning/

Stanford Human-Centered Artificial Intelligence. (2024, May 2). *Escalation risks from LLMs in military and diplomatic contexts* (Policy brief). https://hai.stanford.edu/policy-brief-escalation-risks-llms-military-and-diplomatic-contexts

United Nations Office for Disarmament Affairs. (2024). *Convention on Certain Conventional Weapons: Group of Governmental Experts on lethal autonomous weapons systems (2024)*. https://meetings.unoda.org/ccw-/convention-on-certain-conventional-weapons-group-of-governmental-experts-on-lethal-autonomous-weapons-systems-2024

U.S. Department of Defense Chief Digital and Artificial Intelligence Office. (n.d.). *CDAO*. Retrieved January 11, 2026, from https://www.ai.mil/what

The Washington Post. (2024, February 20). *Pentagon holds conference on AI and LLM applications in warfare*. https://www.washingtonpost.com/technology/2024/02/20/pentagon-ai-llm-conference/