Technology & National Security Review

VOLUME 1

MIT / HARVARD TECHNOLOGY & NATIONAL SECURITY CONFERENCE

April 3–4, 2026 | MIT Kresge Auditorium & HBS Klarman Hall | Cambridge, MA

Finalist Papers

1. **Architecting Trust: A Modular Framework for the Operational Deployment of Autonomous Systems** (Joel Pederson)…….....5

2. **Frozen, Blind, and Air-Gapped: Reconciling Frontier Model Capabilities with Defense Infrastructure Realities** (Steven Varshavsky and Naveen Krishnan)……...33

3. **Detecting Systematic Infrastructure Attacks via Geospatial Intelligence** (Kevin Chen and Cole Griffiths)……....70

4. **When Bans Don't Build Markets: Rethinking the DoD's Critical Mineral Strategy** (Bethany Russell)…….….90

5. **The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities** (Daria, Amy Chang, Erich Devendorf, Michael Kouremetis, Tiffany Saade)……121

6. **De-Risking Defense Innovation at the Earliest Stages: The Strategic Role of Entrepreneurial Fellowships and Early Non-Dilutive Grant Funding** (Elizabeth Kennedy and Lauren Emmi)……...…162

7. **Why Workforce Governance Is the Limiting Factor in National Security Innovation** (Desiree Lorell)……...199

8. **Escalation Protocols for Autonomous Naval Vessels: A Legal-Technical Framework for Safe and Credible Maritime Autonomy** (Jordan Foley)……225

Inaugural Review Committee

MIT / Harvard Technology & National Security Conference

LEAD REVIEWERS

Jonathan Qu	*MBA, Massachusetts Institute of Technology*
Bethany Russell	*MBA/MPP, Harvard University*

FACULTY REVIEWERS

The following faculty judges selected the top three papers from the finalist pool.

Jake Sullivan	*Kissinger Professor, Harvard Kennedy School; Former National Security Advisor*
Kathleen Hicks	*Senior Fellow, Belfer Center, Harvard Kennedy School; Former Deputy Secretary of Defense*
Eric Evans	*Former Director of MIT Lincoln Laboratory, now Director Emeritus, with decades of leadership in advancing U.S. defense technologies*

STUDENT REVIEWERS

Andrea Howard	*PhD Candidate, Massachusetts Institute of Technology*
Steven Varshavsky	*MPP, Harvard Kennedy School*
Naveen Krishan	*MPP, Harvard University; MBA, University of Pennsylvania (Wharton)*
Connor Elkin	*MS/MBA, Harvard University*
Aidan Kenealy	*MBA, Massachusetts Institute of Technology*
Cait Toole	*Harvard Law School*
Dan Tapia	*MBA, Massachusetts Institute of Technology*
James Sorenson	*MBA, Harvard University*
Alex Santangelo	*MPP/MBA, Harvard University*

Cam Heard *MBA Fellow, Massachusetts Institute of Technology*

EVALUATION CRITERIA

Novelty — 25%

Veracity / Academic Rigor — 25%

Impact on Defense — 50%

A NOTE ON THE REVIEW PROCESS

To ensure fairness of grading, committee members did not review their own papers. All other papers were assessed by the full review committee.

Architecting Trust: A Modular Framework for the Operational Deployment of Autonomous Systems

Joel Pederson

System Design and Management, Massachusetts Institute of Technology

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Note

The views expressed in this paper are those of the author and do not reflect the official policy or position of the author's employer.

Correspondence concerning this article should be addressed to Joel Pederson, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Building E40-315, Cambridge, MA 02139.

Email: pederson@mit.edu

Author Bio

Joel Pederson is a Senior Systems Engineer specializing in complex algorithm development within the defense sector. He is currently pursuing an M.S. in Engineering and Management through the MIT System Design and Management (SDM) program. He holds an M.S. in Engineering Management from Northeastern University and a B.S. in Mechanical Engineering from the University of Hawaii at Manoa.

Abstract

For the United States and its allies, the incorporation of non-deterministic Artificial Intelligence (AI) and Machine Learning (ML) systems into tactical platforms presents significant challenges. Certifying these agents for safety-critical operations remains a considerable barrier to their deployment and widespread adoption. This paper examines how AI-enhanced tactical solutions can be effectively fielded despite the inherent risks involved. The analysis begins by discussing the history of automation bias in weapons systems and the novel vulnerabilities introduced by AI/ML-powered solutions. The challenge with these emerging techniques is that unlike traditional software, these algorithms present risks throughout their lifecycle. From adversarial perturbations that can be introduced during model training to stochastic failures during operation, the risks associated with AI/ML-based algorithms cannot simply be mitigated by the legacy safety systems embedded in platforms today. While traditional "physics-based" guardrails (e.g., automated ground collision avoidance) effectively prevent kinematic disasters, they on their own lack the sophistication to address the cognitive and perception errors inherent to modern AI. To safely proliferate AI/ML solutions, a new safety-focused reference architecture is required.

This paper proposes using a modular "Safety Sidecar" architecture that operates across the software's lifecycle. The research defines a strategy that acquisition authorities can consider building from to systematically embed safety barriers directly into the system's training, perception, and planning loops. The framework outlines how the training of models can be protected and refined for operational fielding through a structured lifecycle assurance approach. Specifically, the architecture introduces a "Perception Gatekeeper" which validates input integrity against adversarial or degraded sensor data in real-time. The framework also integrates algorithms running operationally in the loop with AI/ML models to function as a "Model

Constraint Guardian" to enforce safety barriers directly into the system's perception and planning loops. The Safety Sidecar concept can function as a unified framework for AI/ML lifecycle assurance. The approach can facilitate effective integration of safety into both developmental and operational phases of system development. This contribution provides a structured pathway to help ensure that safety is a continuous property from model training through to battlefield deployment.

Keywords: AI Assurance, Trusted Autonomy, Runtime Assurance, Autonomous Weapons Systems, Safety-Critical AI

Architecting Trust: A Modular Framework for the Operational Deployment of Autonomous Systems

The continual evolution of warfare is reaching the limits of human cognitive capacity. Offensive capabilities are becoming more affordable, intelligent, and interconnected, while defensive platforms are increasingly automated and decentralized. This shift represents a fundamental change in the nature of conflict. Warfare is rapidly evolving across all domains from subsea to space, driven by technological proliferation[1]. Over the past 90 years, computing has increasingly made its way into warfare as a means of enhancing lethality through augmenting standalone human capability via automation.

This trajectory was broadly anticipated by early visionaries. Vannevar Bush's 1945 essay "As We May Think"[2] described the concept of a machine that could assist humans with the process of thinking while remaining under direct control by the operator. Decades later, Steve Jobs famously described the computer as "a bicycle for our minds"[3]. However, neither Bush nor Jobs could have completely foreseen the profound leap in computing from that of a passive assistant to an autonomous agent. Today, basic machine learning algorithms, which are used as small components within a larger software architecture such as those supporting threat assessment on the Navy's Aegis combat system[4], require human input in order to take an action. In these legacy paradigms, the software recommends an engagement, but a human operator remains "in the loop" to verify the decision and prevent fratricide.

While this "human-in-the-loop" safety architecture has served relatively well in the past, it simply cannot keep pace with the speed at which future conflicts will be fought. As offensive drone swarms and hypersonic missile threats compress decision timelines, the human operator becomes the primary bottleneck in the kill chain. It is clear that increased autonomy will be

needed to augment human warfighting capabilities[5,6,7], but blind trust in a system's autonomous capability has risks. History has shown that erroneous decisions made by tactical software can lead to fratricide and mission failure[8].

The pace of warfare's evolution combined with rapid developments in commercial Artificial Intelligence (AI) and Machine Learning (ML) present both risks and opportunities for the United States (U.S.) and its allies. In order to maintain asymmetric advantages over adversaries, autonomous systems must be deployed across domains from subsea to space. However, the Department of Defense (DoD) faces a certification issue. Traditional software safety standards, such as DO-178C, rely on deterministic logic where every line of code is verifiable. Modern AI tools, commonly powered by deep neural networks, are non-deterministic "black boxes". They are susceptible to hallucinations, adversarial perturbations, and stochastic failures that traditional "physics-based" guardrails cannot comprehensively detect.

Placing trust in non-deterministic AI-powered black boxes rightfully gives U.S. and allied forces pause. Policymakers hesitate to advocate for proliferated autonomy when the risks are uncertain. Commanders cannot trust an agent with lethal autonomy when it remains so fundamentally vulnerable to taking the wrong action based on sensor noise or adversarial spoofing. Simultaneously, the industrial base faces barriers in delivering capabilities that cannot be verified by traditional means. Yet the strategic necessity of autonomy remains. We cannot wait for a "perfect" AI solution. Holding out for a capability that is 100% explainable and error-free is a strategy for obsolescence.

Instead, modular "Safety Sidecars" should be plugged into autonomous systems to validate inputs and outputs in real-time. This paper describes a reference architecture that decouples "safety" from "intelligence."

Importantly, this is a conceptual framework rather than a validated solution. This modular reference architecture is presented as a starting point to structure the conversation among systems engineers, researchers, and acquisition professionals about how safety can be systematically embedded into AI/ML systems. Substantial research, prototyping, and testing will be required before these concepts can be deployed on tactical platforms.

The proposed architecture builds upon the emerging Runtime Assurance (RTA) architecture (ASTM F3269-17) which, although initially developed for unmanned aircraft systems, contains principles applicable to autonomous systems more broadly. However, the standard requires a more robust framework to be applied to a broader set of missions. Current implementations utilize simple physics-based monitors that cannot detect the semantic failures (e.g., hallucinations) inherent to modern AI. This conceptual architecture suggests a pathway by which acquisition programs could certify safety constraints without needing to fully decipher the model. This is accomplished by implementing an enhanced RTA framework referred to as the "Model Constraint Guardian", which serves as the operational safety module. The Guardian consists of two integrated components: a "Perception Gatekeeper" to validate sensor data and an "Action Governor" to enforce operational bounds.

Background and Motivation

The Limits of Current Safety Paradigms

The application of autonomy in military platforms is not new[9]. However, systems to date have relied on direct human oversight. They operate with limited scope to take actions based on their perception of the threat environment. Machine Learning algorithms powered by Neural Networks or Bayes Classifiers have been focus areas of military research for decades and are now being integrated into fielded systems[10,11,12,13]. For example, these techniques now support threat assessment on the Navy's Aegis combat system[4]. In this paradigm, the tactical software takes in sensor data from radars on the ship and runs it through machine learning algorithms to recommend an engagement decision. Importantly, a human operator remains in the loop to assess the decision made by the system's algorithms[14]. This adds a layer of safety through human verification to avoid fratricide, though it is not a perfect solution. Erroneous decisions made by tactical software have resulted in lethal consequences in both active conflict and peacetime[8,14].

The perpetual evolution of warfare is compressing kill chains to the point where keeping a human in the loop to review engagement decisions may soon no longer be reasonably feasible. Hypersonic weapons, drone swarms, and other emerging threats are designed to create dilemmas for defenders by reducing response times and overwhelming decision makers. Military strategists view AI/ML-infused systems as the necessary solution to close this reaction-time gap. Yet the traditional protections used to ensure that fielded autonomous capabilities are only employed as authorized simply are not built for the next generation of AI/ML-powered platforms. While traditional physics-based guardrails help systems effectively prevent kinematic disasters[15], they

on their own lack the sophistication to address the cognitive and perception errors inherent to modern AI solutions.

Novel Vulnerabilities in the Kill Chain

The integration of AI/ML techniques precipitates a paradigm shift in system failure modes. Traditional tactical software is fragile but predictable — failing only when explicit logic is violated. ML techniques are capable but inscrutable. Neural networks, for example, are forged through large-scale optimization across millions of parameters, creating a statistical model that forms its own internal logic. While this allows for complex behaviors, it obfuscates verifiable cause-and-effect relationships through probabilistic correlations, rendering traditional low-level requirements-based verification methods obsolete.

The vulnerabilities of AI/ML solutions span the entire system lifecycle, from model development through operational deployment. During the development phase, the training data supply chain presents a critical attack vector. As highlighted by the AI data security information document[16] released by the National Security Agency and some Five Eyes partners, the "data supply chain" (p. 1) is a primary national security risk. "Maliciously modified" (p. 1) data can be injected into foundation models that are then acquired and fine-tuned by defense contractors. This data can corrupt the ML-model learning process in potentially highly targeted ways, leading to off-nominal behaviors which will directly lower the model's performance.

During operational deployment, tactical platforms face attacks across the sense-decide-act cycle that can cascade through the entire kill chain. In the sensing phase, adversarial inputs can manipulate system perception. For example, recent research has demonstrated that specialized physical stickers attached to targets can trick unmanned aerial vehicles (UAVs) into failing to correctly classify targets[17]. Researchers were able to achieve an average attack success

rate of 82.0% and maintained a high level of effectiveness when transferring between different detection models. Another paper exposed vulnerabilities in the digital domain, specifically within Support Vector Machine (SVM) defenses designed to detect GPS spoofing[18]. Researchers introduced two adaptive strategies: a data location shift attack, where the vehicle's GPS position slowly drifts to avoid immediate detection, as well as a similarity-based noise attack which hides malicious deviations inside patterns that mimic nominal GPS static. Through their simulations they showed that they could successfully trick the protection algorithm causing its accuracy to fall from 99.9% to as low as 20.4%.

These corrupted inputs then cascade through decision logic and manifest as catastrophic action failures. Research from Purdue demonstrated how imperceptible noise in data inputs can fundamentally corrupt an agent's decision-making logic[19]. High-value strategic moves in their simulation were replaced by irrational or even self-destructive commands. The AI directed erratic physical behaviors such as oscillating in place or targeting empty space, that degraded performance in highly obvious and operationally fatal ways.

The "Valley of Death" for Safety Assurance

Operational AI safety verification for defense systems is in its infancy. Academic research in the field is robust, yet the capabilities required for operational deployment remain immature. An analysis of the current research landscape indicates that leading algorithmic candidates for safety verification are broadly at Technology Readiness Level (TRL) 3. Two primary examples from laboratories working on safe autonomy in complex worlds at the Massachusetts Institute of Technology (MIT) illustrate this maturity gap.

The Constraint-Aware Refinement for Verification (CARV) algorithm introduces a novel method of efficiently verifying whether a model's action space remains within defined safe

bounds[20]. The Certifiable Algorithm for Shape estimation and Tracking (CAST#) produces mathematical proof in the form of a certificate that the object pose and shape estimation is the globally optimal solution based on the sensor input provided and model-based object matching[21]. Together, these algorithms demonstrate AI safety solutions that can be integrated into the training process (CARV) and operate in real-time during operation to keep AI algorithms within safe bounds (CAST#).

Both algorithms demonstrate the promise of AI safety but lack technical maturity. These capabilities function in controlled, highly specific environments but have not yet been integrated into larger software architectures. The real challenge lies in the integration of algorithms like CARV and CAST# into a variety of new and legacy weapon systems.

Just as there are inherent risks in adopting AI-infused technologies into military applications, there is inherent risk within AI safety solutions as well. The acquisition community perceives these novel, unproven safety technologies as an additional risk rather than a risk mitigator due to their immaturity and lack of operational validation. To mature this technology beyond TRL 3, a new architectural approach is required. Safety algorithms must be integrated into end-to-end software systems, allowing them to be validated via Hardware-in-the-Loop (HWIL) testing. By utilizing a Modular Open Systems Approach (MOSA) to host these algorithms, it is possible to bridge the "Valley of Death" for the deployment of tactical AI safety capabilities.

The "Safety Sidecar" Reference Architecture

Decoupling Intelligence from Assurance

To address the trade-off between the rapid evolution of AI models and the rigid requirements of safety-critical functions in military systems, this paper proposes adopting a "Safety Sidecar" reference architecture. A sidecar is a secondary software container that runs in parallel with the primary application. The concept is common in the commercial software world. In this paradigm, data logging or security encryption exist within the sidecar while the main program focuses on its core functionality. The approach allows developers to standardize and update supporting functions without modifying the core functionality.

While "sidecars" are traditionally viewed solely as runtime monitors in commercial software (e.g., Kubernetes), for autonomous defense systems, the concept must extend beyond runtime and into the development and training lifecycle. The framework is not a novel invention of this research, but rather an advocacy for the broader adoption of the Sidecar Pattern as applied to defense systems.

This architecture posits that "Trust" is not a single operational check but a continuous chain of custody which begins with training inputs and continues through to tactical operation. The architecture consists of two distinct but synchronized environments:

1. *The Developmental Guardian (Design-Time):* Located within the training and simulation environment, this layer utilizes computationally intensive formal verification algorithms (such as CARV) to mathematically validate that the AI agent's action space remains within safe bounds. Because this occurs offline, it is not constrained by the size, weight, and power (SWaP) limits of the tactical edge. Its role is to certify the "Safe Action Space" of the model before it is ever compiled for deployment.

2. *The Operational Sidecar (Run-Time):* Upon deployment, a lightweight, deterministic instance of the sidecar travels with the agent on the tactical platform. Instead of recalculating complex proofs, this module acts as a high-speed "governor." It utilizes real-time monitoring algorithms (such as CAST#) to validate sensor integrity and enforces the pre-validated safety constraints derived during the developmental phase.

The Developmental Guardian — Data Quality and Training

AI/ML model safety begins with foundational data integrity. The integrity of training data is the first vulnerability in the model lifecycle. Before a model can be trained, its curriculum must be secured. As part of the Developmental Guardian concept, strict data provenance protocols must be followed. While data cleaning and preparation are already a fundamental part of the ML pipeline, in a defense context, the provenance of the source must be rigorously traced and vetted before training. Scanning training datasets for statistical anomalies is critical to identifying data "poisoning" attacks or adversarial perturbations. But scanning alone may not reveal hidden vulnerabilities. Data provided by trusted DoD or Intel Community partners can be treated with a far higher degree of trust than commercial sources, which require further vetting to ensure the foundation model is not learning from maliciously modified inputs designed to create dormant vulnerabilities within the system's capabilities.

Once data integrity is established, the training process itself must be bound. While various ML paradigms exist for developing tactical models, Reinforcement Learning (RL) has emerged as a particularly promising technique. This method creates agents that, through training, can discover novel strategies through trial and error. RL offers considerable opportunities for defense applications but also introduces novel risks[22,23]. Traditional RL allows agents to explore actions within a defined "action space" to maximize their "reward". This paradigm can lead to

unsafe behaviors if poorly bounded, as the agent itself has no built-in concept of safety or ethics. To mitigate this, the training pipeline must employ techniques that embed safety constraints. Computational methods, such as CARV, can accelerate the verification of Neural Feedback Loops (NFLs) by computing a conservative "set" of future states and then only refining the calculation when a potential violation is detected. This efficiency allows developers to integrate mathematical safety checks directly into the training loop without stalling the learning process, ensuring the final model is trained to satisfy safety constraints before deployment.

The Operational Sidecar — Runtime Assurance

While the Developmental Guardian helps ensure models are trained safely, the operational environment will continue to present AI systems with novel, unique risks that cannot be fully simulated, let alone envisioned during development. Adversarial attacks such as sensor spoofing or sudden environmental anomalies require real-time deterministic safeguards to protect AI systems. The Operational Sidecar provides continuous assurance for both data coming in and out of the system through its Model Constraint Guardian, which consists of two components: a "Perception Gatekeeper" and an "Action Governor."

The first line of defense addresses the "Garbage In" problem. Neural networks are vulnerable to adversarial perturbations, which can cause them to misclassify inputs with high confidence, leading to false detections—confidently identifying objects that do not physically exist or misinterpreting sensor noise as valid obstacles[24].

To counter this, the Sidecar employs a Perception Gatekeeper. Unlike the primary AI, which focuses on perceptual anomalies (e.g., "Target Identified"), the Gatekeeper focuses on metric integrity. It utilizes certifiable algorithms to validate that the physical state of the world is mathematically consistent with sensor readings. Algorithms like CAST# (described earlier) can

detect phantom tracks by verifying that sensor measurements support a physically consistent object—filtering out hallucinated threats or adversarially projected decoys that lack geometric validity. This ensures the vehicle only reacts to physically verified objects.

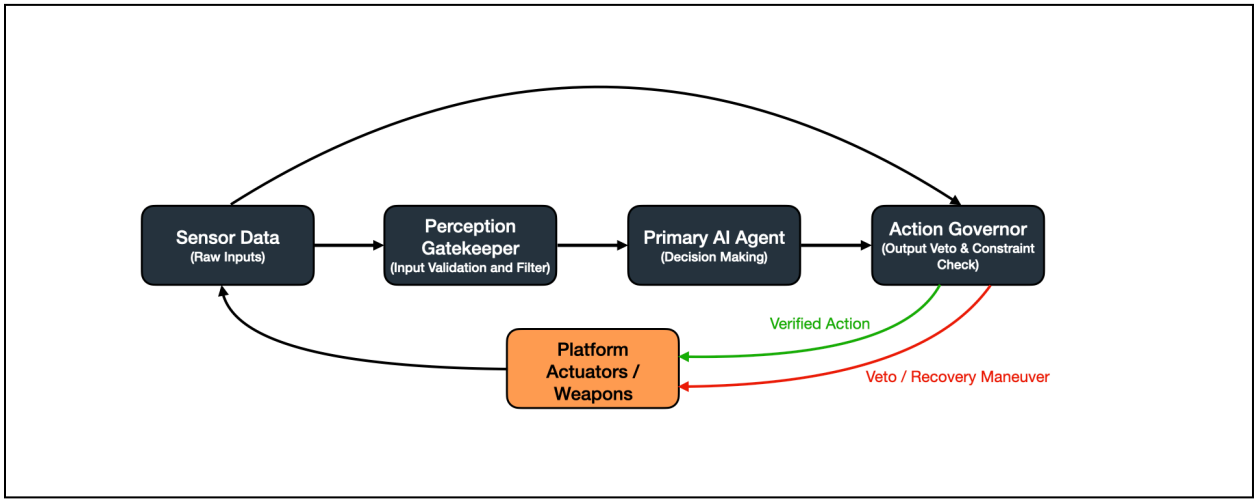
The final layer of defense addresses the "Garbage Out" problem. Even with valid inputs, a stochastic model such as a neural network may occasionally propose unsafe maneuvers due to distribution shift.

This risk is mitigated by the Action Governor, which acts as a deterministic "Veto" layer between the AI agent and the system's output surfaces (e.g., actuators, engagement command, etc). This mechanism aligns with the concept of safety "shielding" discussed in formal methods[25] and reinforcement learning literature[26]. In this framework, the protection is layered. The Perception Gatekeeper (discussed above) handles the semantic failures that traditional control cannot see (e.g., preventing an attack on a misclassified object). The Action Governor then handles the kinetic failures that traditional control excels at (e.g., preventing ground collision).

By combining these two layers, the architecture addresses a critical coverage gap: the Semantic layer catches the "hallucinations" (which physics-based guards miss), while the Kinetic layer catches the "stochastic glitches" (which might cause a crash). If the AI proposes a command that violates either the semantic rules of engagement or the physical laws of flight, the

Governor overrides the signal and engages a deterministic recovery maneuver.

Figure 1
Operational Control Loop with Sidecar Implementation



This architecture bridges the gap between probabilistic AI and deterministic Systems Engineering. Because the Operational Sidecar utilizes rule-based logic to enforce the final "veto," it allows the system to satisfy traditional "Shall" statements (e.g., 'The system shall not engage targets within 500m of a No-Strike List entity'). While the neural network agent itself cannot be formally verified against such absolute requirements due to its stochastic nature, the Sidecar can be. This decoupling allows the verification strategy to focus on the safety wrapper rather than the opaque neural network, a concept detailed further in Section 5.

Modular Independence

A critical feature of the Sidecar architecture is the strict decoupling of the AI agent from the safety monitor. The Safety Monitor, hosted on an isolated compute partition (e.g., a separate core or an independent microcontroller), can act as a Watchdog.

This separation implements the Run-Time Assurance (RTA) architecture defined in industry standard ASTM F3269-17 and detailed in NASA's framework for autonomous operations[27]. The system gains two strategic advantages from the approach:

1. *Asynchronous Evolution:* The AI model can be updated to quickly evolve with the adversary's changing tactics. The lengthy recertification process will not be necessary for model updates as long as the Safety Sidecar remains stable and unchanged. This allows the system to stay relevant on the battlefield without triggering a full recertification cycle for every software update.
2. *Fault Isolation:* A crash, memory leak, or infinite loop in the complex AI container cannot cause the safety monitor to freeze. If a critical fault occurs in the system, the Sidecar can execute pre-programmed recovery maneuvers (e.g., Return-to-Base). The sidecar possesses the control authority to save the system despite lacking the intelligence to complete the mission.

Concept of Operations

A notional defensive stress test scenario is described using the Collaborative Combat Aircraft (CCA) to illustrate the operational benefit of the proposed architecture. This vignette illustrates that relying on precise, deterministic state data as safety monitors do today is insufficient when operational constraints (such as Emissions Control) degrade sensor fidelity. The scenario highlights the need for an architectural layer that enforces safety through uncertainty management rather than static state-based rules.

A formation consisting of one manned 5th-Generation Fighter (Lead) and one CCA wingman is operating in contested airspace, maneuvering to evade a hostile Surface-to-Air Missile (SAM) threat. In order to preserve the formation's low-observable signature, the mission requires strict stealth discipline. The CCAs are prohibited from using active radar to track the lead as a result. The formation enters a dense cloud layer while maneuvering. The environmental

conditions blind the passive Electro-Optical (EO) sensors, while simultaneous hostile jamming degrades data link capability, leading to a 1 second latency in message traffic.

While in this degraded state, the Human Lead executes a high-G maneuver to break the SAM lock. At the same time, the data link dropout is depriving the CCA of real-time telemetry for the Lead. The primary AI agent for the CCAs, a deep reinforcement learning policy trained to prioritize formation integrity in this mission context, falls victim to a model overconfidence error. The CCA's neural network lacks up-to-date sensor data and must rely on internal state-prediction models. These models erroneously extrapolate the Lead's previous straight-and-level state vector. Because the AI is optimizing for formation station-keeping, the AI commands a turn to re-join the predicted position of the Lead. In reality, this command places one of the drones on a collision course with the maneuvering fighter.

In a standard architecture, the safety monitor would validate this catastrophic command. The monitor evaluates the safety of the maneuver based on the reported state of the system, determining that the perceived separation distance is greater than 1,000 feet and that the relative velocity vector appears to be divergent. Because the standard monitor checks the model's confidence rather than the data's quality and context, it fails to detect that the source of truth is flawed. The system validates the CCA's turn command, leading to a mid-air collision between the autonomous wingman and manned asset.

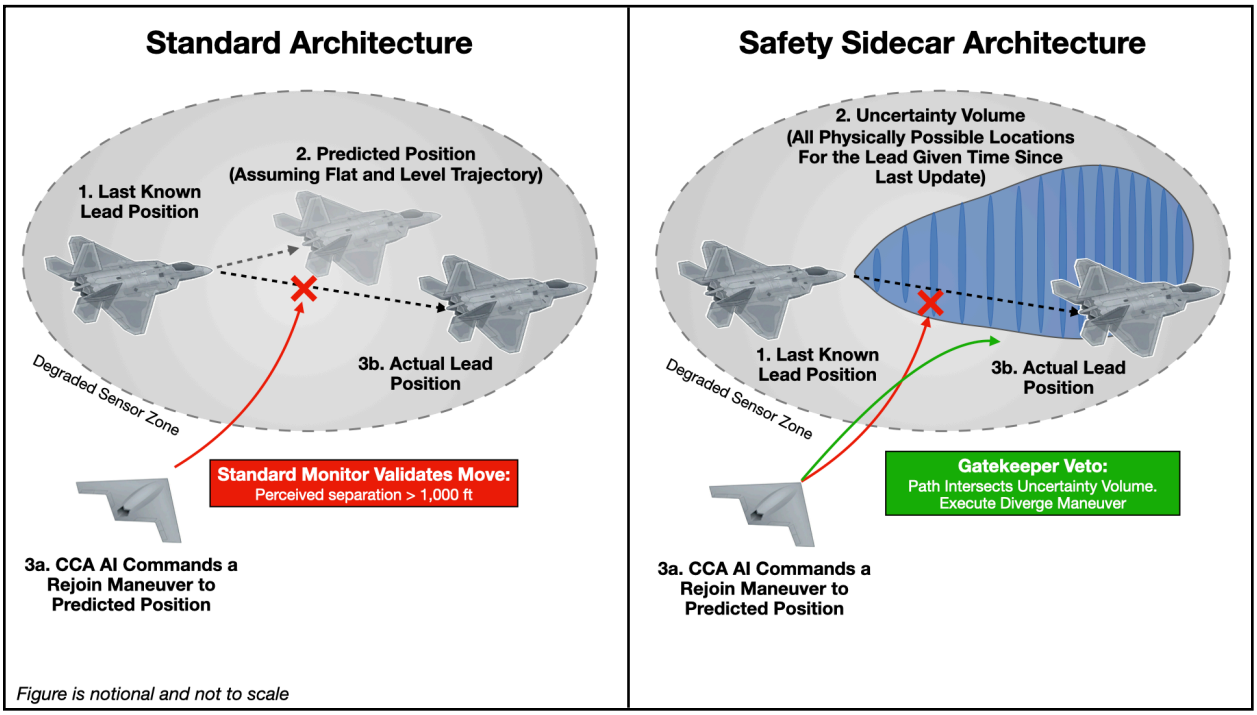
In the proposed architecture, the Perception Gatekeeper would intervene by accounting for the epistemic uncertainty of the state. Instead of validating a point-in-time snapshot, the architecture would leverage control barrier functions based on the age of the data. The perception gatekeeper would not ask where the Lead is, but rather where it is physically possible for a maneuvering fighter to be after one second of silence. This calculation would generate a rapidly

expanding "Volume of Potential Presence", a probabilistic danger volume around the Lead's last known position.

Upon detecting that the AI's proposed flight path intersects this expanded volume, the Gatekeeper would veto the rejoin maneuver. The Action Governor would override the primary agent and force a deterministic "Lost Visual / Lost Link" protocol, executing a lateral divergence maneuver. This response would guarantee physical separation by respecting the bounds of the unknown, rather than the false certainty of the AI model. This vignette demonstrates that safety in autonomous systems is an epistemic constraint; the architecture must be capable of dynamically expanding safety buffers in direct proportion to information degradation.

Figure 2

CONOPS Comparing Standard and Safety Sidecar Architectures



Note. A side-by-side comparison of architectures using CCA as a notional example. The left panel depicts a Standard Architecture, where the primary AI commands a rejoin based on a stale, linear state vector, failing to account for the lead fighter's dynamic maneuvering within the degraded sensor zone. The right panel demonstrates the Safety Sidecar Architecture resolving this uncertainty. The Perception Gatekeeper computes a bounding uncertainty volume (reachability set) encompassing all physically possible locations of the lead fighter given the time since the last update. Upon detecting that the AI's proposed path intersects this strict kinematic boundary, the Action Governor vetoes the unconstrained command and enforces a deterministic divergence maneuver to ensure safe physical separation. The F-22 and CCA aircraft vector assets utilized in this diagram were generated using Google Gemini[28].

Architectural Verification and Implementation

The viability of a parallel, semantic-safety architecture ultimately depends on whether it can be implemented within the rigid constraints of modern DoD systems. This section explains how the "Model Constraint Guardian" architecture can resolve the primary barriers to fielding advanced autonomous systems across the joint force by navigating software certification, managing edge-compute limitations, and ensuring policy compliance.

Certification Through Isolation

One of the fundamental bottlenecks in fielding autonomous combat systems is the safety certification process. Across DoD, software safety is governed by standards such as MIL-STD-882E. This standard lays down rigorous Software Hazard Risk Assessments in order to achieve the highest Levels of Rigor for software whose failure could result in catastrophic loss of the platform, fratricide, or civilian casualties.

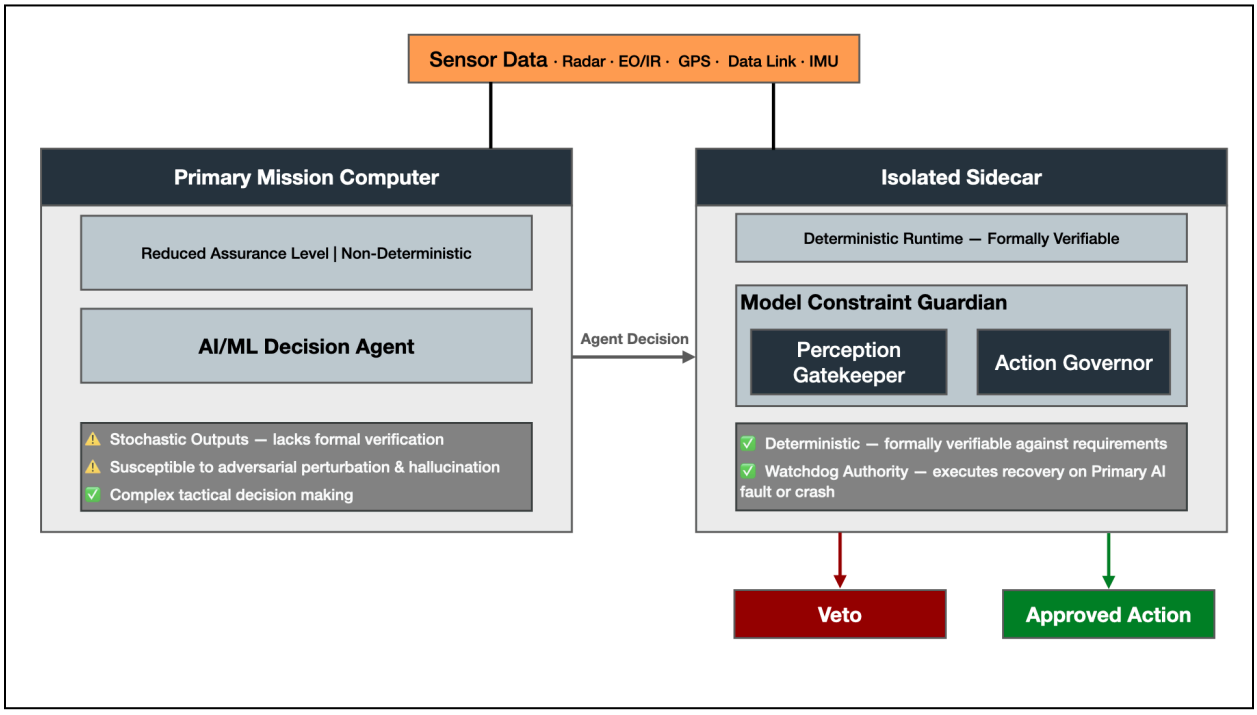
In defense systems today, there are already narrow applications of AI/ML in use to support perception, target classification, and decision support[4]. These applications treat the AI/ML solution as more of an advisory input rather than a safety-critical authority. Deep Reinforcement Learning agents and complex Neural Networks designed for dynamic vehicle control are inherently non-deterministic. Verification of a multi-mission-parameter neural network across all possible state spaces is currently mathematically impossible. As a result, using unconstrained, probabilistic AI as a final authority for lethal action or flight-critical safety effectively bars these systems from achieving the Levels of Rigor required for expanded deployment.

The proposed semantic architecture could resolve this bottleneck through the adoption of a "Simplex Architecture" pattern. By separating the safety-critical constraints from the mission-

optimization logic, the certification burden is isolated. The non-deterministic Primary AI can be evaluated at a lower assurance level, recognizing that its failure is functionally mitigated by an external system. The Model Constraint Guardian relies on deterministic physics equations, bounded reachability sets, and clear rules. These are all able to be formally verified to MIL-STD-882E standards. This architectural isolation forms a pragmatic pathway to safely field advanced AI solutions across the Joint Force, helping to ensure that verifiable safety constraints remain paramount.

Figure 3

Simplex Architecture for Autonomous System Safety Certification



Note. This architecture physically and logically isolates the non-deterministic AI agent from the deterministic Model Constraint Guardian. Containing the complex tactical decision-making within a reduced assurance environment restricts the safety-critical certification burden (e.g., MIL-STD-882E compliance) solely to the formally verifiable sidecar.

Computational Overhead and Edge Deployment

The unique computational burden of running real-time semantic gatekeepers is considerable. Whether deployed on an attributable drone, an Unmanned Surface Vessel, or a Robotic Combat Vehicle, the strict Size, Weight, and Power (SWaP) budgets are a limiting factor.

While the Action Governor avoids the 'curse of dimensionality' by utilizing bounded, short-horizon approximations of reachable sets rather than full predictive simulations, the Perception Gatekeeper's independent semantic verification still requires significant processing throughput. Whether the Guardian is correlating secondary sensor data, running deterministic target-template matching, or cross-referencing complex Rules of Engagement (ROE) logic matrices, it must process high-fidelity data in real-time without introducing latency into the kill chain.

Dedicated, hardware-accelerated compute at the tactical edge is required in order to prevent the Guardian from competing for resources with the Primary AI's processing pipelines. It must be deployed on computer boards that are physically isolated from the primary mission computer. This separation of the hardware provides a physical "air gap" that protects the semantic safety monitor from software faults originating in the Primary AI, as previously discussed.

Policy Compliance and the Digital Proxy

Autonomous weapons systems must comply with stringent legal frameworks in addition to technical safety. Department of Defense Directive (DoDD) 3000.09 mandates that autonomous systems allow commanders to exercise "appropriate levels of human judgment over the use of force."[29].

A monolithic neural network operating as a "black box" cannot demonstrate this compliance, as its internal logic for a specific targeting decision is largely opaque. The Model Constraint Guardian serves as the literal, digital embodiment of the commander's intent. By transforming Rules of Engagement (ROE) and the Law of Armed Conflict (LOAC) into explicit, geometric constraints hosted within the Guardian (e.g., "target must possess a geometrically verified weapon system, not just a probabilistic visual resemblance"), the architecture provides a transparent, auditable mechanism for compliance. The Sidecar acts as a deterministic method to ensure that the lethal bounds established by human judgment are structurally enforced, regardless of the emergent tactical choices made by the primary agent.

By acknowledging that modern AI/ML solutions will remain inherently probabilistic, the defense acquisition community can avoid the impossible task of perfecting the algorithm and instead focus on architecting deterministic bounds around it. This transition from algorithm-centric verification to architecture-centric safety is the foundational step required to move tactical AI further onto the battlefield.

Recommendations for Future Work and Test & Evaluation

While the proposed modular architecture provides a structural pathway for safely fielding autonomous systems, the realization of this capability at an operational scale requires further research across human-machine teaming, coalition interoperability, and legal frameworks. The DoD operates a large variety of systems across a multitude of domains and contexts. A one-size-fits-all architecture is unlikely to be universally adopted. Future work must build upon the architectural foundation to refine it for specific system implementation. Transitioning from theory to implementation will require resolving the simplifications and assumptions inherent in

this conceptual framework, ultimately yielding tailored architectures that are operationally trusted by the Joint Force and its allies, and can be more widely adopted.

"Trust Calibration" associated with human factors is the first major hurdle to address for autonomous architectures. While the Safety Sidecar structurally bounds the AI, the human command authority must still understand and, more importantly, trust those bounds. If a Perception Gatekeeper or Action Governor is tuned too conservatively, it may frequently veto the Primary AI, leading to "alert fatigue" or operator frustration and endangering reaction times. Conversely, if it is too permissive, it risks catastrophe. Future studies will be needed to investigate how to effectively communicate the Sidecar's deterministic boundaries to human operators through advanced human-machine interfaces. Commanders need to be able to maintain appropriate calibrated trust in the system without falling into the trap of automation bias.

The defense community must also explore how this modular safety framework translates to coalition warfare and allied interoperability. If U.S. platforms utilize one set of deterministic safety constraints while allied autonomous systems use another, it creates operational friction and fratricide risk. Future research should focus on establishing shared, international standards for Runtime Assurance mechanisms on specific programs of high priority to both the U.S. and select allies.

Finally, future policy work must focus on the continuous evolution of frameworks like Department of Defense Directive (DoDD) 3000.09. While the Sidecar provides a mechanism to enforce Rules of Engagement (ROE) through geometric constraints, empowering systems to operate with expanded operational authority will require clear, defined legal frameworks. Future initiatives will require multidisciplinary teams of systems engineers, operational commanders,

and military legal counsel to develop standardized methodologies for encoding the Law of Armed Conflict into systems with increasing levels of autonomy.

Conclusion

The nature of warfare's continual evolution and the proliferation of autonomous threats place the traditional human-in-the-loop tactical architectures at serious risk. To maintain asymmetric advantage, the United States and its allies must deploy AI-enhanced autonomous systems across all warfighting domains safely and responsibly. Inherent vulnerabilities in non-deterministic deep learning models, governed solely by legacy physics-based safety paradigms, create an unacceptable level of operational and certification risk, ranging from data poisoning to semantic hallucinations.

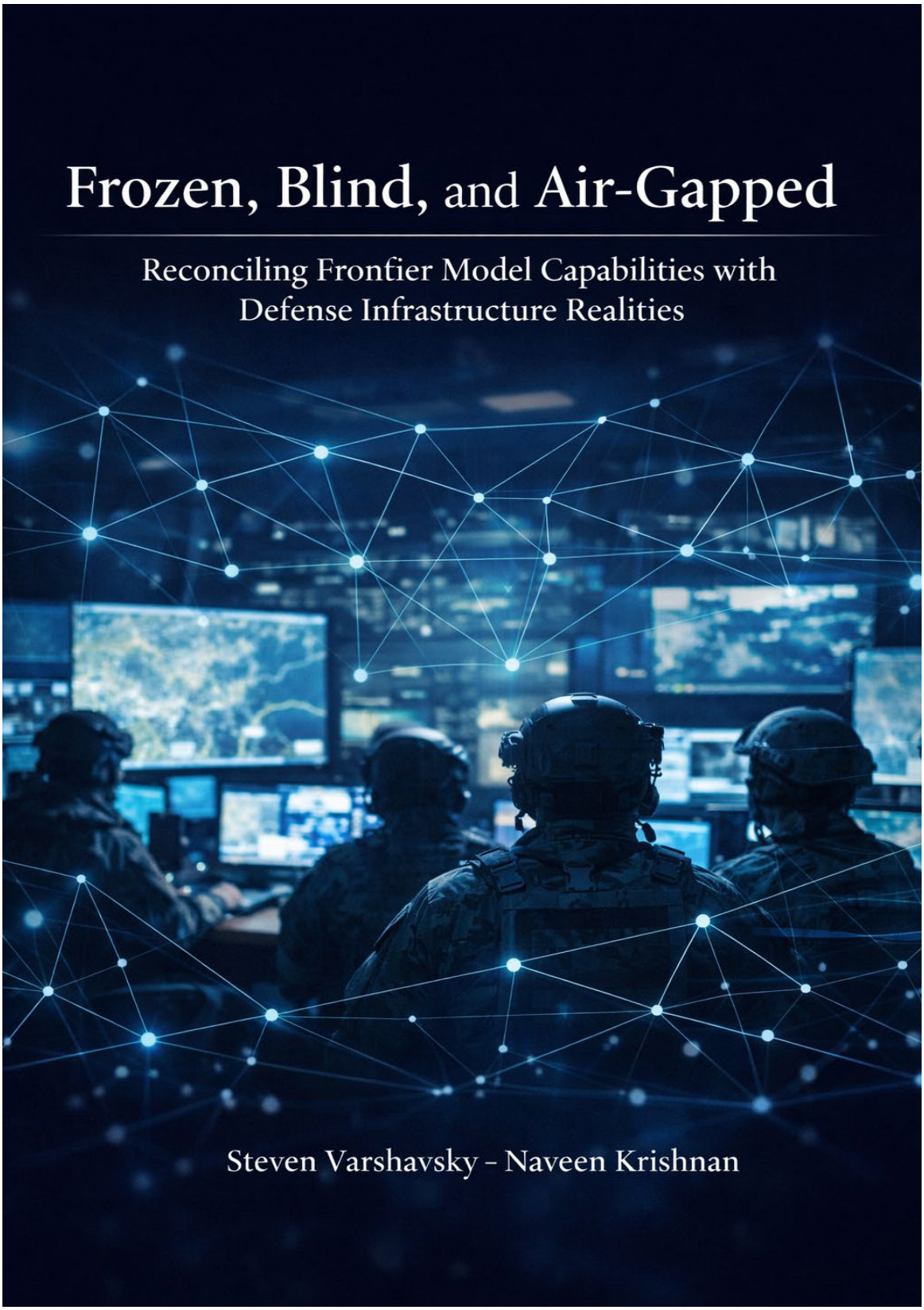
The defense community cannot afford to wait for the invention of perfectly explainable Artificial Intelligence. To field these capabilities in near-term conflicts, trust must be architected. By implementing a modular "Safety Sidecar" framework, programs can establish a continuous chain of safety custody throughout the software lifecycle. During model training, the developmental guardian can ensure models are mathematically bound before deployment. During operation, the Perception Gatekeeper and Action Governor, contained within the Operational Sidecar, provide real-time, deterministic veto authority required for safe, predictable execution at the tactical edge.

By physically and logically isolating the nondeterministic "intelligence" from verifiable "safety", this architecture provides a structural framework to satisfy the rigid computational and policy constraints of modern defense engineering. Architecting trust through modular isolation provides a pragmatic and scalable path forward to field the next generation of autonomous capabilities.

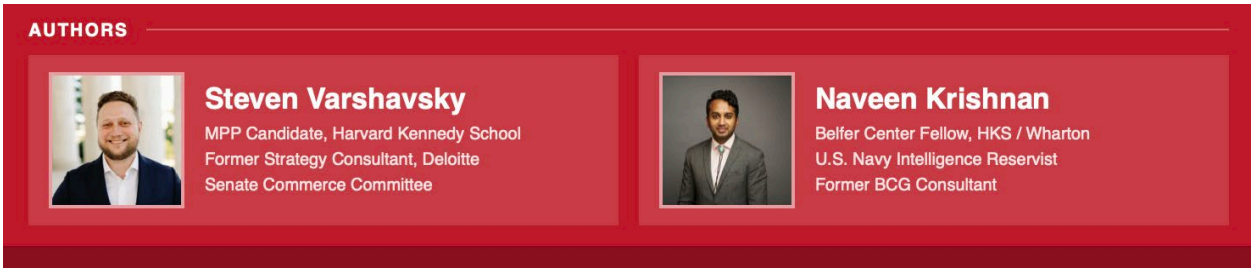
References

1. Pusztaszeri, A., & Harding, E. (2025, September 16). Technological evolution on the battlefield. *CSIS*. https://www.csis.org/analysis/chapter-9-technological-evolution-battlefield
2. Bush, V. (1945, July). As we may think. The Atlantic. https://www.theatlantic.com/magazine/archive/1945/07/as-we-may-think/303881/
3. Lawrence, M. R. (Director). (1990). Memory and imagination [Film]. Library of Congress.
4. Lockheed Martin. (2024, July 29). Lockheed Martin leverages AI and machine learning to revolutionize defense and space technology. https://www.lockheedmartin.com/en-us/news/features/2024/lockheed-martin-leverages-ai-and-machine-learning-to-revolutionize-defense-and-space-technology.html
5. Burdette, Z., Phillips, D., Heim, J., Geist, E., Frelinger, D., Heitzenrater, C., & Mueller, K. P. (2026, January 22). How artificial intelligence could reshape four essential competitions in future warfare | RAND. https://www.rand.org/pubs/research_reports/RRA4316-1.html
6. Rascoe, A. (2025, August 3). How are drones changing what it means to wage war?. *NPR*. https://www.npr.org/2025/08/03/nx-s1-5487561/how-are-drones-changing-what-it-means-to-wage-war
7. Schmidt, E. (2025, August 12). The dawn of automated warfare. Foreign Affairs. https://www.foreignaffairs.com/russia/dawn-automated-warfare
8. Talbot, D. (2005, June 1). Preventing “fratricide.” *MIT Technology Review*. https://www.technologyreview.com/2005/06/01/230882/preventing-fratricide/
9. Zequeira, M. (2024, September). Artificial Intelligence as a combat multiplier. Army University Press. https://www.armyupress.army.mil/Journals/Military-Review/Online-Exclusive/2024-OLE/AI-Combat-Multiplier/
10. Kovaliv, O., Kondratenko, Y., Shevchenko, A., Sidenko, I., & Kondratenko, G. (2023). Neural network architectures for recognizing military objects on satellite images. 2023 IEEE 12th International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 175–180.
11. Cannon, C. T., & Goericke, S. (2021). Using convolution neural networks to develop robust combat behaviors through reinforcement learning [Master’s thesis, Naval Postgraduate School]. NPS Archive: Calhoun.
12. Lippmann, R. P. (1994). Neural Networks, Bayesian a posteriori Probabilities, and Pattern Classification. In From Statistics to Neural Networks : Theory and Pattern Recognition Applications (pp. 83–104). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-79119-2
13. MIT Lincoln Laboratory. (1988). DARPA neural network study (DTIC Accession No. ADA207580). Defense Technical Information Center. https://apps.dtic.mil/sti/citations/ADA207580
14. Singer, P. W. (2009, January 28). In the Loop? Armed Robots and the Future of War. Brookings Institution. https://www.brookings.edu/articles/in-the-loop-armed-robots-and-the-future-of-war/
15. National Aeronautics and Space Administration. (2023, September 6). Automatic collision avoidance technology. https://www.nasa.gov/reference/auto-gcas/

16. National Security Agency, Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, Australian Signals Directorate's Australian Cyber Security Centre, New Zealand Government Communications Security Bureau National Cyber Security Centre, National Cyber Security Centre United Kingdom. (2025, May 22). AI data security: Best practices for securing data used to train & operate AI systems. https://media.defense.gov/2025/May/22/2003720601/-1/-1/0/CSI_AI_DATA_SECURITY.PDF
17. Xi, H., Ru, L., Tian, J., Lu, B., Hu, S., Wang, W., & Luan, X. (2025). URAdv: A novel framework for generating ultra-robust adversarial patches against UAV object detection. Mathematics, 13(4), 591. https://doi.org/10.3390/math13040591
18. An, S., Jang, D. J., & Lee, E.-K. (2025). Adversarial evasion attacks on SVM-based GPS spoofing detection systems. Sensors, 25(19), 6062. https://doi.org/10.3390/s25196062
19. Dabholkar, A., Hare, J. Z., Mittrick, M., Richardson, J., Waytowich, N., Narayanan, P., & Bagchi, S. (2024). Adversarial attacks on reinforcement learning agents for Command and control. The Journal of Defense Modeling and Simulation: Applications, Methodology, Technology, 23(1), 177–190. https://doi.org/10.1177/15485129241271178
20. Rober, N., & How, J. P. (2024). Constraint-aware refinement for safety verification of neural feedback loops. IEEE Control Systems Letters, 8, 3219–3224. https://doi.org/10.1109/lcsys.2024.3518912
21. Shaikewitz, L., Ubellacker, S., & Carlone, L. (2024). A certifiable algorithm for simultaneous shape estimation and object tracking. IEEE Robotics and Automation Letters, 9(12), 11873–11880. https://doi.org/10.1109/lra.2024.3501684
22. Qu, P., Liu, H., Xu, S., Yu, T., Chen, Y., Wang, E., & Na, L. (2025). *Multi-agent deep reinforcement learning for cooperative path planning of UAV swarms.* Research Square. https://doi.org/10.21203/rs.3.rs-6508231/v1
23. Bunch, K., Hou, A. C., Haberman, R., Herron, M., Jacques, A., & Briggs, G. J. (2024). *Risk Assessment of Reinforcement Learning AI Systems: Looking Beyond the Technology.* (Report No. RR-A1473-1). RAND Corporation. https://doi.org/10.7249/RRA1473-1
24. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR). https://arxiv.org/abs/1412.6572
25. Bloem, R., Könighofer, B., Könighofer, R., & Wang, C. (2015). Shield synthesis: Runtime enforcement for reactive systems. In C. Baier & C. Tinelli (Eds.), Tools and Algorithms for the Construction and Analysis of Systems (pp. 533–548). Springer, Cham. https://doi.org/10.1007/978-3-662-46681-0_51
26. Alshiekh, M., Bloem, R., Ehlers, R., Könighofer, B., Niekum, S., & Topcu, U. (2018). Safe reinforcement learning via shielding. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2669–2678. https://doi.org/10.1609/aaai.v32i1.11797
27. Skoog, M., & Hook, L. (2020). Leveraging ASTM Industry Standard F3269-17. NASA Technical Reports Server. https://ntrs.nasa.gov/api/citations/20200001821/downloads/20200001821.pdf
28. Google. (2026). Gemini (Advanced) [Large language model]. https://gemini.google.com
29. Office of the Under Secretary of Defense for Policy. (2023). Autonomy in weapon systems: DOD directive 3000.09. Department of Defense.


Frozen, Blind, and Air-Gapped
Reconciling Frontier Model Capabilities with Defense Infrastructure Realities
Steven Varshavsky - Naveen Krishnan

About the Authors



Steven Varshavsky is a Master in Public Policy candidate at Harvard Kennedy School and a researcher focused on defense innovation and emerging technology. He previously advised U.S. Air Force leaders as a strategy consultant at Deloitte and worked on the Senate Commerce Committee on technology and oversight issues.

Naveen Krishnan is a Belfer Center (BYL) Fellow and dual-degree MPP/MBA candidate at Harvard Kennedy School and Wharton. He is a U.S. Navy Intelligence reservist who researches critical technologies and geopolitics across the Indo-Pacific region. He previously researched US-China relations as a Boren Fellow in Taiwan and Liu Xiaobo Fellow to the US Congressional-Executive Commission on China. Before graduate school, he worked as a consultant with Boston Consulting Group.

EXECUTIVE SUMMARY · 2026 HBS–MIT TECH & NAT SEC CONFERENCE

Frozen, Blind, and Air-Gapped

Reconciling Frontier Model Capabilities with Defense Infrastructure Realities · Varshavsky & Krishnan

> *"The warfighter* **needs the model** *that is allowed to think, in real time, where it matters most (not necessarily what the latest model is.) Building the infrastructure to deliver this model (repeatedly and at scale) is the defining challenge of defense AI for the next decade."*

6–18	560	$33.6K	300W	3×
Months for IL6 ATO	Staff hours per authorization	Cost per ATO cycle	F-35 ICP max power	Compute tiers Cloud→IL6→Edge

1 The Problem: A Structural "Security Tax"

The U.S. fields the world's most advanced AI models yet cannot deploy them where operational advantage is decided. The barrier (instead of model capability, commercial investment, or political will) is a structural misalignment between how frontier AI is built and how defense systems must operate. Three dimensions define the tax:

AUTHORIZATION LATENCY

IL6 ATO takes 6–18 months. Models are frozen at certification while commercial providers ship multiple new generations. DoD perpetually deploys yesterday's AI.

COMPUTE DIFFERENTIAL

The F-35 ICP runs at ~300W max — frontier inference is categorically impossible. Compute shrinks by orders of magnitude at each tier: cloud → IL6 → tactical device.

DATA FRAGMENTATION

Decades of siloed programs and inconsistent metadata leave models "blind" inside air-gapped IL6 systems. RAG pipelines lack the coherence commercial AI assumes.

2 Case Evidence & Model Typology

Project Maven, GIDE, and Agile Flag 24-3 each demonstrate the same finding: AI succeeds when model design, data architecture, and infrastructure are aligned — and stalls when they are not. Success at the tactical edge required bounded mission scope, curated data, and edge-native compute. Systems that assumed centralized cloud reachback failed under DDIL conditions.

> *"Capability that cannot be certified, powered, and updated at the point of need is useless."*

Not all use cases require frontier models. The correct design center maps across environment (CONUS cloud vs. tactical edge) and task scope (broad reasoning vs. narrow domain):

- **Frontier models** suit enterprise functions in permissive, well-connected environments — e.g., GenAI.mil at IL5.
- **SLMs are superior** — not merely acceptable — for ISR, logistics, and cyber defense at the tactical edge.
- **Hybrid tiered architectures** bridge the two: SLMs locally, larger models for rear-echelon synthesis.

3 Policy Roadmap: Three Lines of Effort

1	2	3
Modular IL6-Ready Platform	**Defense Hardware Pipeline**	**Sustained RDT&E as Program of Record**
A pre-certified government runtime where SLMs deploy with incremental re-authorization. Multi-tier topology with automated drift detection and async updates. Extends SWFT & cATO to edge environments.	DPA Title III purchase commitments for low-SWaP inference chips. OSC loan guarantees ($984M authority) for edge AI firms. OTA procurement to bypass traditional acquisition timelines. Precedent: Title III scaled GaN MMIC chips for radar.	Dedicated NDAA line items for reference SLM development, curated datasets, and secure update pipelines. FY2025 NDAA §§1532–1533 provide footholds. CDAO as integrating authority.

4 Conclusion: Infrastructure as Advantage

The DoD's current posture is one of **importation** — forcing commercial models into defense infrastructure. This works in permissive environments and fails structurally at the classified and tactical edge. A fine-tuned SLM on a low-SWaP IL6-certified device delivers more combat power than a frontier model in a CONUS cloud enclave, regardless of benchmark scores.

Sustainable AI advantage comes from **reproducible infrastructure**. Maven, GIDE, and Agile Flag all succeeded as pilots but never scaled because the certification pathways, hardware pipelines, and data architecture to replicate them didn't exist. Allied nations — NATO DIANA, UK MoD, Dstl — independently arrive at the same answer: small, certifiable, edge-viable models. The mismatch is infrastructural, not algorithmic. **The tools and authorities exist. The task is alignment.**

Introduction

As artificial intelligence (AI) advances, its integration into military systems is reshaping the future of warfare. It offers the potential to equip warfighters with decisive advantages: enabling faster, more informed decisions through real-time connectivity and data-driven insight. Turning that promise into operational reality depends on overcoming deep structural barriers in how the military procures and integrates emerging technologies. The core problem lies in the incompatibility between security standards, built for the advent of modern cybersecurity, and the acquisition architecture required to diffuse cutting edge models throughout the warfighting tech stack.

Functionally, the Department of Defense has already begun to leverage AI in various capacities including predictive maintenance, logistics optimization, and intelligence, surveillance, & reconnaissance (ISR) missions. For instance, predictive maintenance uses AI algorithms to analyze data from aircraft sensors, predicting potential failures before they occur. Logistics optimization employs AI to streamline supply chains, ensuring that troops receive the necessary resources efficiently and effectively.[i] This capability, which currently exists only in certain pockets of innovation, if proliferated across each major weapon system, would provide the U.S. military with actionable real-time insight ensuring no equipment would undergo unnecessary improvement, or experience unplanned maintenance.

Project Maven, operating under the DoD's Joint Artificial Intelligence Center (JAIC) is the U.S. military's cornerstone attempt to leverage machine learning in order to augment existing object detection capabilities. The Maven Smart System fuses various training and operational data sources: satellite imagery, geolocation data, and communications intercepts, into a unified interface for battlefield analysis.[ii] Eventually individual users, rather than subcontractors, will be able to create their own applications and interfaces for their individual use cases, (instead of relying on a complex and arduous subcontracting process.) The long tail of this story arc is all data, available to cleared users, for any purpose.

Predictably, AI is most useful in data available and data rich environments because it can learn from and leverage both static and dynamic human-in-the-loop workflows. Even if doctrine could

adequately catch up with tactical capability delivery, the result would still be a military fighting with yesterday's AI, constrained by its own authorization architecture instead of an adversary. Denied, Degraded, Intermittent, and Low-Bandwidth (DDIL) conditions are the predicted operating baseline in a contested Indo-Pacific theater, and the actual operating baseline today in Ukraine. Combined Joint All-Domain Command and Control (CJADC2), the Department of Defense's concept for linking sensors, shooters, and decision-makers across domains into a unified network, assumes rapid data exchange across distributed platforms.[iii] Yet the very environments in which CJADC2 must function are those in which persistent connectivity, high-bandwidth reach back, and centralized cloud compute cannot be guaranteed.

This mismatch is structural: Frontier AI models are architected around continuous access to hyperscale infrastructure, dense telemetry, and rapid update pipelines. Warfighting systems must instead function under intermittent communications, constrained spectrum, and size-, weight-, and power (SWaP)-limited hardware. When reachback fails, inference must occur locally. When bandwidth collapses, data must already be curated and staged. When oversight cannot travel at the speed of fiber, decision-support logic must already reside at the edge.

Diagnosis

The United States now confronts a paradox in the application of artificial intelligence in national security: it fields the world's most advanced language and reasoning models yet struggles to deploy them where operational advantage is decided. This paper argues that a growing "security tax" constrains military AI, not imposed by adversaries or law, but by the structural misalignment between frontier model design and the Department of Defense's classified and edge computing environments. By the time a model is accredited for Impact Level 6 (IL6), it is effectively frozen in time: disconnected from live data, severed from hyperscale cloud compute, and governed by security code that evolves far slower than AI code.

Frontier labs that contract with the DoD cannot, inherently, use their full powered, commercially available data center infrastructure in highly classified DoD cloud environments. Top Secret information upon which DoD models must be trained requires IL6 classification: an arduous certification process requiring dedicated air-gapped facilities, network isolation, and access

controls that precipitate a roughly 12 month Authorization to Operate (ATO) process.[iv] This notwithstanding the length of time it takes to develop powerful enough compute capacity to operate in these environments, or the fact that each system, device, platform, and weapon system that operates on an IL6 approved environment requires separate configuration.[v] It should come as no surprise that the recently announced "GenAI.mil" platform[vi] operates via Google Gemini, given that only a few months ago, Google achieved IL6 certification both for its distributed cloud undergirding Gemini and Vertex AI systems, and air gapped appliances.[vii] This code cannot be changed without compromising security, but it was designed for static systems meant to house databases and discrete applications, and not dynamic AI models with weekly updates. Security architecture (IL5/IL6) therefore represents the single biggest impediment to deploying frontier AI in military systems, not a lack of models or political will. The security "code" is too slow for the AI "code."

Using Intelligence, Surveillance, and Reconnaissance (ISR) and Combined Joint All-Domain Command and Control (CJADC2) as core cases, the paper compares idealized commercial AI/MLOps architectures with current IL5/IL6 and tactical-edge stacks to expose where this security tax accumulates. Frontier models assume continuous, high-bandwidth access to centralized cloud resources, rich observability, and rapid update cycles, while warfighting systems must operate in denied, degraded, intermittent, and low-bandwidth (DDIL) conditions on SWaP-constrained platforms.

CJADC2 combined with cutting edge frontier models, as promised by Project Maven, would deliver real time data fusion across military services and amongst partners, AI-enabled decision support, and edge computing on weapon platforms such as ships, jets, bombers, and drones. Exercises such as Global Information Dominance Experiments and Agile Flag autonomous logistics demonstrations show that AI can deliver real value at the edge, but only when models, data, and compute are explicitly designed for that environment. During Bamboo Eagle/Agile Flag exercises in August 2024, an Air Expeditionary Wing demonstrated semi-autonomous airlift capability to deliver over 20 urgently needed Mission Capable Parts orders to multiple geographically separated locations: displaying the promise and practical use of AI capabilities at the tactical edge.[viii]

Our analysis diagnoses three interlocking constraints: compute, freshness, and data hygiene.

First, edge platforms cannot host full-scale frontier models without unacceptable SWaP and latency penalties. To operate tactical edge models, they cannot, by definition, connect back to the CONUS compute infrastructure, and therefore must be brought entirely along to the tactical edge. For reference, A concrete anchor is the Integrated Core Processor (ICP) fielded on the F-35, which L3Harris lists at ~36 lb with ~300 W (max) power dissipation for a short (1 ATR) configuration (already a meaningful thermal and electrical load inside a tightly packed, cooled avionics environment.)[ix] Therefore, edge SWaP requirements force operators into using smaller models and/or heavier compression, and materially lower general capability should be expected under these conditions unless narrow tasking, retrieval, or domain fine-tuning can be employed to compensate.

Second, the IL6 authorization and Authorization to Operate (ATO) process lags frontier model update cadence, creating "frozen" models that are outdated on arrival. These exercises also uncovered and articulated IL6 constraints to operations with tactical edge AI stemming primarily from the need to operate individualized modular tech stacks at the tactical edge. Models must be small and portable to collect, analyze, and process data, and must be certified for classified deployment. Further, each platform, whether its an F-35, drone, or missile system, requires a separate configuration to the tactical edge AI.

Third, decades of fragmented, unstructured defense data is accessed and leveraged through blunt retrieval-augmented generation, leaving even powerful models "blind" inside air-gapped systems. Much of the Department's operational data remains siloed across legacy systems, inconsistently labeled, access-restricted by program office boundaries, or stored in formats never designed for machine-readable integration. Even where data technically exists, it is often not curated, normalized, or authorized for cross-domain retrieval at the classification level where the model operates. The Government Accountability Office reported that overly restrictive data classification was directly responsible for the interoperability of data used for command and control purposes.[x] The result is that retrieval pipelines inside IL6 environments lack the density, cleanliness, and semantic coherence that commercial models assume.

Rather than forcing massive commercial models into architectures they were never designed to inhabit, the paper proposes a strategic pivot: invest in small, mission-specific language models trained on curated internal datasets, deployed on low-SWaP tactical hardware, and supported by a pre-certified, reusable IL6 model platform. Policy tools such as Defense Production Act Title III and targeted RDT&E remediations through NDAA line items can underwrite reference SLMs, edge accelerators, and secure update pipelines, creating a repeatable bridge from commercial innovation to certified combat capability. This framework reframes "AI advantage" as infrastructure alignment, arguing that warfighters need a model allowed to think, in real time, where it matters most over the 'largest' one.[xi]

Literature Review/Background

The Department of Defense's engagement with artificial intelligence has followed a decade-long trend from experimentation to institutional integration, which has included organizational restructuring and persistent scaling failures. The following background explains this trajectory & the structural barriers that have accompanied it to provide context for the security tax framework this paper later develops.

I. Evolution of DoD AI Integration (2016–2025)

The modern era of DoD AI begins with Project Maven, initiated in early 2017 as the military's first large-scale attempt to apply machine learning, specifically computer vision, to operational intelligence workflows.[xii] Maven operated under the Joint Artificial Intelligence Center (JAIC), established in June 2018 under the DoD Chief Information Officer to coordinate department-wide AI efforts and accelerate delivery of AI-enabled capabilities. However by 2021 the institutional architecture had already proven insufficient. The JAIC, Defense Digital Service, Office of the Chief Data Officer, and the Advana analytics program were consolidated into the Chief Digital and Artificial Intelligence Office (CDAO), which became fully operational in June 2022. The CDAO's November 2023 Data, Analytics, and Artificial Intelligence Adoption Strategy unified prior guidance, emphasizing federated data infrastructure, responsible AI adoption, and warfighter decision advantage as first-order priorities. Most recently, the

Department of War's January 2026 AI-first memoranda under Secretary Hegseth directed department-wide access to frontier generative AI models and mandated component-level identification of "pace-setting projects" within 30 days, *signaling an executive-level shift from experimentation toward operational integration.*[xiii] Yet the pattern persists: pilot programs succeed in controlled settings, but scaling across the enterprise stalls at the boundaries of certification, classification, and hardware constraints.

II. Compute and Edge Infrastructure Constraints

The gap between what *frontier* AI requires and what *tactical* environments can provide is fundamentally a compute problem. Commercial large language models are architected for hyperscale data centers with elastic GPU clusters, high-bandwidth interconnects, and virtually unlimited power. Warfighting systems operate under the inverse conditions.

- Denied, Degraded, Intermittent, and Low-bandwidth (DDIL) environments are the predicted operating baseline in a contested Indo-Pacific theater and the actual baseline in Ukraine today.
- Edge platforms impose strict SWaP constraints that preclude hosting full-scale frontier models.

As one defense industry assessment noted, military teams require "high-performance computing in a rugged, SWaP-optimized form factor" that is fundamentally incompatible with commercial off-the-shelf AI infrastructure.[xiv] DARPA's multi-billion dollar AI Next campaign and the OFFSET program for autonomous swarm systems both reflected early an recognition: edge-viable AI demanded purpose-built architectures (instead of scaled-down commercial ones.) More recently the CDAO's Edge Data Mesh (EDM), which was successfully demonstrated during Global Information Dominance Experiment 13 and Project Convergence Capstone 5 in April 2025, represents the most concrete attempt to solve tactical data distribution in DDIL conditions. Notably, CDAO officials left EDM nodes in place with Indo-Pacific Command for operational use after the exercise, moving the Department closer to "bi-directional, real-time data flow between the tactical edge and operational and strategic decision-makers".[xv]

To make the Edge Data Mesh operational rather than aspirational, the Department should implement it as an API-centric federation of mission data products, instead of as another

centralized data lake with spokes at the tactical edge. Under the Department of the Air Force's API Reference Architecture, authoritative systems should expose data through registered, versioned APIs in an enterprise catalog, with API gateways handling request validation, routing, and coarse-grained access enforcement so that edge and enterprise consumers can discover and consume the same data products through standardized interfaces. Zero Trust and Identity, Credentialing, and Access Management (ICAM) then become the control plane for the mesh: data should be tagged at creation, mapped to user, device, and contextual attributes, and released through policy decision and enforcement points that enable dynamic, attribute-based access rather than static network-based trust. In practice, this would mean that each IL6 enclave, tactical cloud, or forward node hosts only the mission-relevant data products it needs, synchronizes them asynchronously when connectivity permits, and relies on federated identity to preserve auditability and least privilege across Air Force, joint, and mission-partner environments. The result is a data mesh that can support edge-deployed SLMs with clean, discoverable, policy-governed data access without requiring the Department to collapse its entire data estate into a single repository.[xvi]

III. Security and Authorization Bottlenecks

The Authorization to Operate (ATO) process is the single largest procedural impediment to deploying current-generation AI in classified environments. Traditional ATO requires a static, point-in-time security assessment, typically renewed every three years, consuming approximately 560 hours of manual effort from a team of four assessors at an estimated cost of $33,600 per assessment.[xvii] For Impact Level 6 environments, which handle classified Secret data, the requirements grow: IL6 exceeds FedRAMP High baselines, demanding dedicated air-gapped facilities, full network isolation, and the complete complement of NIST 800-53 controls. A full ATO at this level can take six to eighteen months, during which time frontier model providers may release multiple major model updates that the authorized version will never incorporate.

The DoD's Software Modernization Implementation Plan for FY25-26 acknowledges this tension directly, calling for continuous ATO (cATO) frameworks that grant ongoing authorization based on demonstrated cybersecurity maturity rather than periodic reassessment.[xviii] The FY2025 NDAA included provisions to reduce ATO barriers, and the Pentagon's Software Fast Track

(SWFT) program aims to replace legacy authorization with automated, continuous monitoring. These reforms are promising but nascent; the vast majority of classified AI deployments *still* operate under the traditional ATO regime, leaving authorized models frozen at the version that survived the certification gauntlet.

IV. Data Hygiene and Fragmentation Across Defense Networks

Even if compute and authorization constraints were resolved tomorrow, deployed models would still face a data problem. Decades of siloed acquisition programs, inconsistent metadata standards, and program-office-level access restrictions have produced a defense data environment that is structurally hostile to modern AI techniques. The Government Accountability Office reported in 2025 that "overly restrictive data classification" directly hindered the interoperability of data used for command and control purposes, identifying it as a critical challenge to achieving CJADC2 objectives.[xix] The CDAO has responded with its federated data catalog initiative, publishing Data Decrees mandating that all DoD data be treated as an enterprise resource, published in the catalog with common interface specifications, and exchanged via automated interfaces. The 2023 CDAO data mesh request for information also signaled intent to implement a zero-trust, cross-classification data architecture for the CJADC2 data integration layer. As Scale AI's Dan Tadross, formerly of the Marine Corps Warfighting Laboratory and JAIC, described the environment: "All the data is in different databases. It's not clean. You don't have the context behind why that data was there".[xx] Retrieval-augmented generation (RAG) pipelines inside IL6 environments therefore lack the density, cleanliness, and coherence that commercial models assume, making even powerful models effectively blind to the institutional knowledge they are supposed to leverage.

V. Comparative Civil-Military AI Governance and Allied Approaches

The structural mismatch this paper identifies is not uniquely American, but the United States' position as the world's leading AI producer makes it the most consequential. Commercial frontier-model pipelines operate on cadences of weeks: rapid iteration, automated evaluation, continuous deployment via self-hosted APIs. However, defense AI MLOps require IL5/IL6 manual validation, separate enclaves, and human-in-the-loop reviews at every stage.[xxi] Allied

nations have begun exploring alternative approaches that prioritize certifiable, lightweight models over frontier-scale capability.

U.S. Allies: NATO's Defence Innovation Accelerator for the North Atlantic (DIANA), operational since 2023, launched its Rapid Adoption Service with an explicit goal endorsed at the 2025 Hague Summit: adopt new technologies and integrate them into Allied armed forces within 24 months.[xxii] One case is the UK Ministry of Defence's engagement with DIANA innovator MyLanguage Inc. to deploy secure, offline voice-to-voice translation supporting 23 languages, eliminating dependence on network connectivity and human interpreters in conflict zones. The UK's Defence Science and Technology Laboratory (Dstl) conducted its largest-ever multi-domain AI trial in partnership with the United States and Australia, collecting data to develop AI systems that reduce cognitive burden while keeping humans central to decision-making. These allied efforts share a common design center: small, certifiable, edge-viable models built for the environments where they will actually operate, rather than frontier models adapted, often unsuccessfully, after the fact.

U.S. Adversaries: China's military-civil fusion facilitates the translation of commercial technology into military applications in ways that spare them from the frozen authorization gap discussed in this paper. Chip export controls have inadvertently forced investment in smaller, domestically certifiable models more suited to bespoke use cases, inclusive of DDIL operations planning by the People's Liberation Army (PLA).

Based on the United States' own assessment of the PLA's progress in AI: China has narrowed the performance gap through continued investment in AI for ISR, electromagnetic warfare (EW), autonomous vehicles, and automated target recognition.[xxiii] The PLA has leaped forward in performance by leveraging MCF between its traditional commercial and defense industrial bases and by studying Project Maven and DIU's OSINT program specifically.[xxiv] The result of these trends presented itself at the 2024 Zhuhai airshow, where Norinco, a traditional Chinese defense contractor, demonstrated autonomously dispatching drones, modeling the battlefield, tracking targets, devising strike plans, and executing follow-up strikes - with human-in-the-loop only for the final fire command.[xxv] These are the exact edge-autonomous, locally inferencing capabilities the US cannot yet field because of the security tax required to field and test them.

Drawing on over 2,800 AI-related PLA contract award notices from January 2023 through December 2024, CSET finds that while legacy state-owned defense enterprises still lead procurement, an emerging set of nontraditional vendors and research institutions is playing a larger role with AI.[xxvi] This development, if scaled, would frustrate the prevailing U.S. approach to hamstringing Chinese military modernization, given that the majority of these new market entrants are not yet subject to U.S. sanctions. (Note: the PLA's acquisition architecture is not immune to corruption and falsification of capability: *100 Trust*, one of the aforementioned nontraditional vendors, was recently caught by the PLA in faking metrics within its bid for military contracts.[xxvii])

Two broad takeaways emerge from the PLA: First, by way of intense U.S. compute restrictions, China is being already forced to do what this paper recommends that the U.S. do voluntarily: build smaller, domestically certified, edge-viable models on constrained hardware. Second, while China is investing in the selfsame suite of edge, AI-enabled capabilities through non-traditional vendors, it still has its own infrastructure and procurement gaps to close to achieve reliable pipelines of capability delivery.

This international alignment on strategy reinforces the core thesis that the mismatch between commercial AI and defense deployment is infrastructural and procedural as opposed to algorithmic.

Methodology

I. Comparative Reference Architectures

This paper employs a comparative reference architecture methodology to diagnose where security and infrastructure constraints impose a "security tax" on the deployment of artificial intelligence within the Department of Defense. Rather than evaluating AI performance abstractly, the analysis compares the architectural assumptions underlying commercial frontier-model MLOps stacks with the realities of DoD Impact Level 5 (IL5), Impact Level 6 (IL6), and tactical edge environments. This paper relies on the 2024 Verma and Santhanam conception and analysis of the requirements of each of these environments and the areas where they overlap (See

Appendix 1, 2, and 3).

Commercial Architecture: Commercial frontier-model MLOps stacks rely on several conditions for their optimal performance. In 2024, Google scholars wrote on the subject of continuous delivery pipelines through MLOps, and assumed that for data scientists and ML engineers who want to apply DevOps principles to ML systems (MLOps), that those data scientists would have access to scale inputs for plentiful data and cloud elasticity and end-to-end automation and monitoring throughout the development lifecycle.[xxviii] Likewise, Amazon's MLOps framework assumes that machine learning workloads can be fully integrated into modern software delivery pipelines. As AWS describes it, "MLOps is the discipline of integrating ML workloads into release management, Continuous Integration / Continuous Delivery (CI/CD), and operations".[xxix] This framing assumes that model development, validation, and deployment occur within environments that support automated build, test, and release cycles, rather than static accreditation baselines. This is the distinction and tension our paper addresses.

IL5 & IL6: In contrast, IL5 and IL6 environments and the requirements associated with operation in them mandate physically and logically isolated environments, often hosted within U.S.-controlled facilities, and limit access to U.S. citizens or personnel with appropriate vetting. In IL5 and IL6 environments CI/CD automation is constrained by accreditation cycles and manual controls under a risk management framework; compute resources must satisfy stringent locality, vetting, and isolation requirements, limiting elasticity; and telemetry and logging systems must meet additional audit, clearance, and access rules, impeding rapid telemetry-driven feedback loops.[xxx] This paper will examine not only how these "code" or design differences manifest in current operational realities, but also discrete, achievable, and modularized solutions that the DoD could employ to circumvent these challenges without compromising security.

Edge Environments: At the edge, these assumptions break down due to requirements that arise from DDIL operations. The DDIL Requirements dimension in our framework deals with issues which arise when one has to adapt and modify a solution developed with a centralized MLOps in mind to work in a distributed environment with a limited network in place. In short, these conditions require adjustments to select, available application stack layers for select, available

MLOps lifecycle stages to be used for select, appropriate DDIL challenges.[xxxi] This paper will suggest unique, independent MLOps reference architectures according to the explicit DDIL requirements distinct from the assumptions made both in CONUS commercial and IL5 and 6 security environments.

II. ISR and CJADC2

This paper employs a structured case-comparison methodology across operational AI efforts that span intelligence, targeting, logistics, and command-and-control domains. Rather than cataloging programs descriptively, the analysis isolates architectural and environmental variables that shaped performance outcomes. The core question is not whether AI "worked," but under what specific technical and institutional conditions it produced measurable operational advantage.

Project Maven provides the first anchor case. Maven demonstrated that machine learning can materially accelerate image and video analysis in ISR workflows, particularly when data pipelines are curated, labeling regimes are tightly managed, and models are embedded into analyst feedback loops. Its success was model performance and more importantly the integration of inference outputs into operational tasking cycles. Where Maven delivered value, it did so within bounded mission sets, defined data domains, and controlled deployment architectures.

Global Information Dominance Experiments (GIDE) provide a second analytic lens. GIDE focused on integrating multi-source ISR data streams to support time-sensitive decision-making across combatant commands. The experiments highlight the role of data integration, cross-domain synchronization, and compute availability in shaping AI effectiveness. Where connectivity, shared data standards, and compute resources were sufficient, AI-supported decision aids accelerated operational tempo. Where data was fragmented or bandwidth constrained, performance degraded. GIDE illustrates that AI advantage scales with infrastructure alignment.[xxxii]

Agile Flag autonomous logistics demonstrations provide a third case. In this context, AI-supported routing, predictive maintenance, and logistics coordination were tested in distributed operational environments. These exercises expose the friction introduced by DDIL conditions, limited edge compute, and human oversight constraints. Where edge autonomy was carefully

bounded and models were optimized for specific tasks, performance gains were realized. Where systems depended on persistent reach-back to centralized compute, resilience suffered.[xxxiii]

Across these cases, three enabling conditions emerge: First, AI delivered value when models were tightly scoped to defined mission problems with curated data inputs. Second, operational benefit depended on infrastructure compatibility (compute availability, networking reliability, and data standardization.) Third, success correlated with lifecycle integration: feedback loops between operators and developers enabled rapid iteration within constrained domains.

The methodology synthesizes these findings into comparative variables (data curation, compute locality, connectivity, accreditation cadence, and operator integration) and evaluates each case along these axes. This enables a disciplined identification of when AI becomes operationally decisive versus when it becomes brittle.

The conclusion is not that AI AI succeeds when *designed explicitly* for contested environments. The cases demonstrate that operational value emerges when model design, data architecture, and infrastructure constraints are aligned rather than assumed away.

III. Designing a Pathway for Edge-Optimized SLMs

Finally, we synthesize the technical constraints (SWaP, DDIL networking, IL6 and ATO processes) with policy instruments (e.g., DPA Title III, targeted RDT&E) to propose a concrete design and funding pathway for small, mission-specific language models optimized for classified and tactical-edge use.

This final methodological step integrates engineering constraints with acquisition and industrial policy tools. The premise: if the "security tax" is structural, then mitigation requires both architectural redesign and targeted policy intervention.

The technical constraints are defined first. SWaP limitations restrict the feasibility of deploying frontier-scale models on tactical platforms. DDIL networking conditions constrain persistent reach-back to cloud compute and inhibit centralized retraining cycles. IL6 environments and Authorization to Operate (ATO) processes impose latency on model updates and restrict

continuous deployment paradigms. Together, these factors render large, cloud-dependent models operationally brittle at the edge.

Rather than forcing hyperscale architectures into SWaP-constrained systems, this paper proposes a pivot toward small, mission-specific language models (SLMs). These models are trained on curated, domain-specific datasets, optimized for low-SWaP hardware accelerators, and designed for intermittent synchronization with higher-tier compute. The architecture emphasizes modular retraining, version control compatible with accreditation cycles, and bounded inference scopes aligned with operator needs.

The policy synthesis component evaluates how such a shift could be financed and industrialized. Defense Production Act (DPA) Title III authorities provide a mechanism to expand domestic production capacity for secure accelerators, edge compute modules, and classified-ready ML toolchains. Targeted RDT&E line items within the NDAA could underwrite reference SLM development, reusable IL6-certified model platforms, and secure update pipelines. Rather than funding isolated prototypes, these instruments would institutionalize a repeatable bridge between commercial innovation and accredited deployment.

The methodology therefore links constraint to intervention. SWaP limitations justify model downsizing. DDIL networking conditions justify decentralized training capabilities. ATO latency justifies pre-certified, reusable model containers. DPA Title III justifies industrial scaling of secure hardware. RDT&E justifies sustained experimentation and iteration.

The outcome is a concrete design-and-funding pathway: develop mission-bounded SLMs; certify a reusable IL6 deployment platform; invest in edge accelerators; and institutionalize secure, modular update cycles. AI advantage is reframed not as model size, but as architectural fit and institutional alignment.

In short, the synthesis moves from diagnosis to prescription. If the problem is structural misalignment, the solution must be structural redesign supported by deliberate industrial policy.

Output

The prior analysis establishes that the Department of Defense faces three interlocking constraints: (1) compute scarcity, (2) authorization latency, and (3) data fragmentation - that collectively prevent frontier AI from delivering operational value at the classified and tactical edge. The following section does the following: formalizes these constraints into a structured framework, introduces a typology for matching model architectures to mission profiles, and proposes a policy-backed roadmap for building the infrastructure required to field AI where it matters.

I. Formalizing the "Security Tax"

The concept of a "security tax" captures a straightforward reality: every layer of security and compliance architecture the Department imposes on an AI system converts raw model capability into integration overhead. This tax accumulates across three measurable dimensions, each of which widens the gap between what commercial AI can do and what defense-deployed AI is allowed to do.

Authorization latency: A traditional ATO requires approximately 560 hours of manual assessment effort from a four-person team, at an estimated cost of $33,600 per cycle, and the full authorization process for an IL6 system can stretch from six to eighteen months.[xxxiv] During that window, frontier model providers like OpenAI, Anthropic, and Google may release multiple major model iterations, each incorporating architectural improvements, safety patches, and capability expansions that the authorized military version will never see. The model that emerges from ATO is, by definition, the model that entered it: frozen at the version that survived certification. The DoD's own Software Modernization Implementation Plan acknowledges this directly, calling for continuous ATO frameworks that replace point-in-time assessment with ongoing risk determination. But cATO adoption remains limited, and the vast majority of classified systems still operate under traditional authorization cycles. The practical consequence is that the Department is perpetually deploying yesterday's AI into tomorrow's operating environment, a latency that compounds with each model generation.

Compute differential: A commercial frontier model operates on hyperscale infrastructure with elastic GPU provisioning, high-bandwidth interconnects, and effectively unlimited power

budgets. An IL6 enclave operates on dedicated, air-gapped hardware with fixed compute allocations and no ability to burst capacity. The same model that runs in milliseconds on a commercial API endpoint may face unacceptable latency or simply fail to load in a constrained enclave. At the tactical edge, the gap widens further: the F-35's Integrated Core Processor, for instance, operates at roughly 300 watts maximum in a tightly cooled avionics environment (a thermal and electrical budget that categorically excludes frontier-scale inference).[xxxv] Moving from CONUS cloud to IL6 enclave to tactical edge device, available compute shrinks by orders of magnitude at each step, and with it, the class of models that can feasibly operate. The implication is architectural: defense AI systems must be designed for the smallest compute envelope in the deployment chain, not the largest.

Update lag: Commercial MLOps pipelines are designed for continuous deployment. Model weights, safety filters, and retrieval indices are updated on cadences measured in days or weeks. Inside an IL6 environment, every software change triggers a compliance review. Every model update requires re-validation against the full complement of NIST 800-53 controls. The result is that even after initial ATO is granted, the deployed model diverges from its commercial counterpart at a rate determined by the bureaucratic throughput of the authorization system as opposed to the technical readiness of the update. Over a twelve-month ATO cycle, a model may fall multiple generations behind its commercial equivalent. Each major update that is deferred accumulates technical debt: downstream integrations, training data pipelines, and user workflows all become increasingly misaligned with the authorized model version.

The security tax, in aggregate, converts each new model capability into an exponential integration cost: more powerful models demand more compute, require longer authorization, and generate more compliance surface area. This failure is a structural property of the current architecture; a reference comparison illustrates the point. In a commercial deployment, a developer pushes a model update to a cloud endpoint; automated testing validates performance; the update is live within hours. In an IL6 deployment, the same update must be reviewed for classification implications, re-assessed against security controls, validated in a separate enclave, and approved through a chain of authorization officials. The identical technical artifact, subjected to defense security architecture, arrives months later, functionally obsolete. The tax is real, measurable, and compounding.

II. Typology of Model Appropriateness

Not all military AI use cases require frontier-scale models, and recognizing this is the first step toward a viable deployment strategy. We propose a simple typology that distinguishes where frontier models are the correct design center from where small, domain-specific language models (SLMs) are not merely acceptable but superior.

The typology maps along two axes: operational environment and task complexity: Operational environment ranges from CONUS cloud (where persistent connectivity, hyperscale compute, and unclassified or IL5 data access are available) to the tactical edge, where DDIL conditions, SWaP constraints, and IL6 classification requirements dominate. Task complexity ranges from narrow, domain-specific functions (entity extraction, format conversion, sensor-data summarization) to broad, open-ended reasoning tasks (strategic planning support, cross-domain analysis, novel scenario generation).

	CONUS Cloud (Permissive Infrastructure)	**Tactical Edge (DDIL / SWaP-Constrained)**
Broad Complexity (Open-ended reasoning, strategic planning)	**Frontier Models** (e.g., Gemini for Gov)	**Hybrid Architectures** (SLM locally with intermittent cloud sync)
Narrow Complexity (Entity extraction, ISR processing)	**Domain-Specific Models**	**Small Language Models (SLMs)** (Optimized for latency/hardware)

Frontier models occupy the upper-left quadrant: high task complexity, permissive infrastructure. In CONUS or well-connected rear-echelon environments, with IL5 data and reliable cloud reachback, frontier models offer genuine advantages. The GenAI.mil platform, powered initially by Google's Gemini for Government at IL5, represents this use case: enterprise-wide access to generative AI for document summarization, policy drafting, risk assessment, and administrative workflows across approximately three million DoD personnel. For these functions, model scale is a feature because the infrastructure can support it.

SLMs dominate the lower-right quadrant:[xxxvi] narrower tasks, constrained environments. As Defense One reported, the Pentagon's own AI leadership is increasingly oriented toward smaller models: "maybe there is a smaller-parameter model that could run on a laptop or run on an edge device, that will still provide the added benefit of accelerating the operational planning process". OpenAI's head of national security policy, Sasha Baker, described the company's work with national laboratories on models that "feel like a large language model" but operate entirely within a secure perimeter, never reaching beyond it. Scale AI's Dan Tadross framed the approach as "the right model for the right purpose," advocating for smaller-parameter models at the edge rather than transporting data center infrastructure to the first island chain.

Mapping this typology to specific mission profiles isolates the design choices:

- ISR processing at the tactical edge, where a drone or sensor platform must classify imagery, extract entities, or flag anomalies without cloud connectivity, is an SLM problem. The model needs domain-specific fine-tuning on curated intelligence data over general-purpose reasoning breadth.
- Logistics optimization in DDIL conditions, where an Agile Flag-style autonomous airlift system must route deliveries across geographically separated locations, requires small models trained on operational logistics data and deployed on ruggedized edge hardware[xxxvii]
- Cyber defense at the network edge, where models must detect anomalies and respond in real-time without reach back, benefits from lightweight, low-latency inference optimized for network traffic patterns rather than the broad knowledge base of a frontier model.

In each case, the correct design center is *not* the largest available model, but is instead the model explicitly architected for the compute budget, data environment, and latency requirements of the mission.

The intermediate space between these quadrants is is where hybrid architectures apply. A forward-deployed unit might run SLMs locally for immediate decision support while intermittently syncing summarized outputs to a rear-echelon node running a larger model for broader contextual analysis (a pattern consistent with the CDAO's Edge Data Mesh concept, which demonstrated bi-directional data flow between tactical and strategic levels at Project Convergence Capstone 5.)[xxxviii] This is the optimal architecture for environments where connectivity is intermittent & compute is stratified.

The key insight of this typology is negative: the default assumption that frontier capability equals military advantage is wrong. Capability that cannot be deployed, cannot be updated, and cannot access relevant data is irrelevant. The model that delivers decision advantage is the one allowed to operate, in real time, within the infrastructure it actually inhabits.

III. Policy and Design Roadmap for Edge-First AI

Translating this analysis into operational reality requires concurrent investment across three lines of effort: a pre-certified model platform, a hardware accelerator pipeline, and a policy instrument stack that funds and sustains both.

Line of Effort 1: Modular, IL6-Ready SLM Deployment Stack - Rather than subjecting each individual model to a full ATO cycle, the Department should invest in a pre-certified model platform: a hardened, IL6-authorized runtime environment into which new SLMs can be deployed with incremental, not full-cycle, re-authorization. The FY2025 NDAA's provisions for reducing authorization barriers and the Pentagon's SWFT program both move in this direction. What is missing is a reference implementation: a government-owned, open-architecture platform specifically designed to host rotating SLMs at IL6, with standardized interfaces for data ingestion, model evaluation, and secure update delivery. The CDAO's Edge Data Mesh provides a partial model for the data transport layer; the model hosting layer requires equivalent investment.

The modular platform this paper proposes should be designed to address four categories of DDIL requirements identified by Verma and Santhanam (2024) in their IBM Research framework for edge MLOps: the Location Problem, the Automation Problem, the Disruption Problem, and the Adaptation Problem.[xxxix] Each results in a distinct design constraint on the IL6-ready deployment stack and demands a corresponding policy response.

The Location Problem: In centralized environments, component placement is trivial since everything runs in one location. At the tactical edge, components of the ML pipeline must be distributed across multiple tiers: a forward sensor or weapon platform, an intermediate node (such as a shipboard server or forward operating base), and a rear-echelon core site. Verma and Santhanam show that even simple ISR data collection scenarios require deliberate decisions about which MLOps activities (data acquisition, model inference, monitoring) execute at which tier.

Insight: Our policy implication is that the pre-certified IL6 platform cannot be monolithic. It must support modular component placement across at least a two-tier (edge-core) and preferably three-tier (edge-tactical cloud-core) topology, with standardized interfaces that allow mission planners to map pipeline components to available compute at each tier. The CDAO's Edge Data Mesh, which demonstrated bi-directional data flow between tactical and strategic levels at Project Convergence Capstone 5, provides a partial model for the data transport layer between tiers, but the compute and model-hosting layers at each tier require equivalent architectural definition.

The Automation Problem: Commercial MLOps pipelines rely on human operators to handle exceptions: data labeling anomalies, model drift alerts, reconfiguration after failures. At the tactical edge, no such human expertise is reliably available. Verma and Santhanam propose policy-based self-managing systems (rule-based automation that specifies actions under defined conditions) as a replacement for manual intervention.

Insight: For the IL6 deployment stack, this means the platform must embed automated model monitoring, drift detection, and fallback logic that can operate without human-in-the-loop oversight for extended periods. The January 2026 AI-first memoranda's directive for the CDAO to make AI "enablers" available across the Department in real time supports this requirement, but

the memoranda focus on enterprise-level enablement rather than edge-specific autonomous operation.[xl] Policy should mandate that any IL6-certified SLM platform include pre-validated, automated model management capabilities (drift detection, automatic rollback to last-known-good model state, and policy-driven reconfiguration) as a certification requirement instead of an optional feature.

The Disruption Problem: Edge components must continue operating when connectivity to the core service is lost entirely (in a DDIL environment, this is the norm.) Verma and Santhanam recommend local backup components, caching of critical data and model artifacts, asynchronous communication replacing synchronous request-response patterns, and disruption-tolerant networking (DTN) techniques that relay information whenever connectivity is restored.

Insight: For the pre-certified platform, this translates into a specific design requirement: every model deployed at the edge must be accompanied by a pre-staged local inference capability that functions independently of any reach back. The platform should support asynchronous model update delivery (queuing updates at the core and pushing them to edge nodes opportunistically during connectivity windows rather than requiring persistent connections.) The Software Fast Track (SWFT) program, launched by Acting DoD CIO Katie Arrington in May 2025 to replace the legacy ATO/RMF process with AI-enabled continuous compliance workflows, should be explicitly extended to cover these edge-specific asynchronous update scenarios, ensuring that model updates queued during disconnected periods receive expedited compliance validation upon reconnection rather than re-entering the full authorization cycle.[xli]

The Adaptation Problem: Edge devices operate under compute, storage, and power constraints that are categorically different from core data centers. Verma and Santhanam catalog a suite of techniques for adapting AI solutions to these constraints: model size reduction through TOFA (Transfer Once for All) schemes, quantization of model weights, knowledge distillation, matrix decomposition, data compression via core-sets, caching to reduce network bandwidth, and duty-cycling to conserve power.

Insight: For the SLM deployment stack, this is where the architecture intersects most directly with the typology of model appropriateness developed earlier in this paper. The pre-

certified platform should include a standardized model optimization pipeline (a sequence of compression, quantization, and distillation steps) that produces deployment-ready SLM variants for defined SWaP envelopes. Rather than requiring each program office to independently solve the adaptation problem, the Department should certify a library of reference SLM configurations optimized for common tactical hardware profiles (e.g., shipboard server, UAV payload processor, dismounted soldier device), reducing the per-program engineering burden and accelerating time-to-edge.

Line of Effort 2: Defense-Led Hardware Accelerator Pipeline - The Department needs purpose-built, low-SWaP inference hardware designed for the thermal, electrical, and physical constraints of tactical platforms. Rather than relying on broad-based industrial subsidies, the Trump administration's own policy architecture provides three instruments purpose-built for this task.

First, Defense Production Act Title III offers the most direct mechanism. DPA Title III authorizes purchases, purchase commitments, subsidies, and loan guarantees to create, maintain, or expand domestic industrial base capabilities critical to national defense (Conference Board, 2025). President Trump has already invoked DPA Title III aggressively: the March 2025 executive order on critical mineral production delegated Section 303 authorities to the Secretary of Defense to boost domestic production, while the April 2025 executive order extended Title III authorities to shipbuilding and port infrastructure.[xlii,xliii] The same authority should be directed toward edge AI inference accelerators. A DPA Title III purchase commitment for low-SWaP inference chips (optimized for the thermal envelopes of avionics bays, UAV payloads, and forward operating bases) would create the demand signal that incentivizes domestic manufacturers to invest in defense-relevant form factors that no commercial market would otherwise sustain. The Air Force Research Laboratory has prior precedent: a Title III project successfully scaled domestic production of Gallium Nitride (GaN) Monolithic Microwave Integrated Circuits for radar and electronic warfare applications, demonstrating that Title III can industrialize niche semiconductor capabilities when properly targeted.[xliv]

Second, the Office of Strategic Capital (OSC) provides a complementary financing mechanism.[xlv] Established in December 2022 and formally enacted into law through the

FY2024 NDAA, OSC can issue loans and loan guarantees to companies developing critical supply chain technologies needed for national security. OSC opened its application window with authority to lend up to $984 million, explicitly targeting technologies including autonomous systems and advanced computing. OSC financing should be directed toward the emerging ecosystem of edge inference hardware companies (firms producing specialized AI accelerator chips, ruggedized compute modules, and SWaP-optimized ML toolchains) that sit below the scale threshold that attracts traditional defense prime investment, but above the prototype stage where DARPA typically operates.

Third, the administration's broader acquisition transformation provides the procurement pathway. Secretary Hegseth's November 2025 Acquisition Transformation Strategy explicitly prioritizes commercial solutions, rapid prototyping, and the "85 percent solution" philosophy: fielding proven capability quickly rather than pursuing unachievable perfection through years of development.[xlvi] The cancellation of the Joint Capabilities Integration and Development System (JCIDS) and its replacement with streamlined forums that tie funding directly to warfighting priorities removes a bureaucratic layer that historically delayed hardware adoption. For edge AI hardware specifically, this means the Department can use Other Transactions Authority (OTA) agreements and the Rapid Capabilities Office mechanisms (both prioritized in the April 2025 executive order on modernizing defense acquisitions) to contract directly with commercial edge compute manufacturers, bypassing the traditional acquisition cycle that adds years between prototype and production.

Line of Effort 3: Sustained Policy and Budget Commitment - One-off demonstrations do not produce enduring capability. The Department needs dedicated RDT&E line items in annual NDAA authorizations that fund reference SLM development, curated training dataset creation, and secure model update pipelines as a program of record as opposed to an experiment. The FY2025 NDAA's provisions for AI computing infrastructure (Section 1532) and data acquisition budgeting (Section 1533) provide initial footholds.[xlvii] DPA Title III can underwrite purchase commitments for edge accelerator hardware, creating demand signals that incentivize domestic manufacturers to invest in defense-relevant form factors. The CDAO, empowered by the January 2026 AI-first memoranda to direct release of DoD data to cleared users and to make AI enablers available across the Department in real time, should serve as the integrating authority (ensuring

that the model platform, hardware pipeline, and data infrastructure converge into a single, repeatable deployment pathway.)[xlviii] The goal is a reproducible infrastructure: a bridge from commercial innovation to certified, edge-deployed combat capability that can be used repeatedly as models, hardware, and threats evolve.

Conclusion

This paper has argued that the Department of Defense's primary barrier to fielding effective AI is fundamentally a structural misalignment between how frontier AI systems are built and how defense systems are required to operate. The "security tax" (accumulated across authorization latency, compute constraints, and data fragmentation) converts commercial AI capability into integration overhead, producing systems that are frozen on arrival, blind to institutional data, and architecturally incompatible with the environments where operational advantage is decided. We argue that a shortage of model capability, commercial investment, or political will are *not* the primary barriers to fielding effective AI.

I. Strategic Reframe: Infrastructure as Advantage

The central reframe this paper proposes is straightforward: AI advantage in a military context is measured by infrastructure alignment (the degree to which compute locality, update cadence, and data plumbing support real-time inference at the point of decision) as opposed to model scale. The key question is whether any model can operate, with current data, on available hardware, within authorized security boundaries, at the point of need. By this metric, a small language model fine-tuned on curated intelligence data, deployed on a low-SWaP edge device, and operating under a pre-certified IL6 runtime delivers more combat power than a frontier model trapped in a CONUS cloud enclave, no matter how impressive its benchmark scores. (We are not dismissing the deployment of commercial, frontier AI.) For enterprise functions in permissive environments, the GenAI.mil platform and its successors represent a legitimate and valuable deployment of commercial AI at scale.

The argument is against the prevailing assumption that frontier capability automatically translates to operational advantage regardless of deployment context. Capability that cannot be certified, powered, connected, and updated where the warfighter operates is irrelevant.

II. Policy Pivot: From Importing AI to Cultivating It

The Department's current posture toward AI adoption is largely one of importation: identify the most capable commercial models, negotiate access, and attempt to integrate them into defense infrastructure. We acknowledge this approach has produced real value in permissive, well-connected environments. However it has systematically failed at the classified and tactical edge, for the structural reasons this paper documents. Our proposed pivot is deliberate and specific.

First, invest in pre-authorized model platforms, not individual model ATOs. A government-owned, IL6-certified runtime environment with standardized interfaces would allow the Department to swap SLMs in and out with incremental re-authorization rather than full-cycle ATO. The cATO framework, the SWFT program, and FY2025 NDAA provisions all lay groundwork, but none yet delivers a reference implementation for model hosting at IL6. Without this platform, each new model deployment repeats the same multi-month certification process from scratch, ensuring that the Department is perpetually fielding yesterday's AI.

Second, build a defense-led hardware pipeline for edge AI inference. DPA Title III purchase commitments, OSC loan guarantees, and OTA-based rapid procurement collectively provide the authorities and capital to industrialize edge inference accelerators without depending on broad-based industrial subsidies. The administration has already demonstrated willingness to use these tools at scale for critical minerals, shipbuilding, and nuclear energy infrastructure. The same instruments, directed toward low-SWaP AI inference hardware, would create a defense-specific supply chain for chips that can power SLMs in environments where no commercial product is designed to operate (avionics bays, UAV payloads, forward operating bases, and shipboard systems.)

Third, treat curated defense data as a strategic asset with dedicated investment. The CDAO's Data Decrees and federated data catalog initiative establish the policy framework; what is needed is sustained funding for data curation, labeling, normalization, and cross-domain access enablement (treated as infrastructure investment on par with hardware procurement.) Retrieval-augmented generation is only as effective as the data it retrieves. Inside an air-gapped IL6 environment, that data must be pre-staged, semantically coherent, and authorized for the

classification level at which the model operates. This is an ongoing operational requirement that demands dedicated personnel, tools, and budget lines.

III. Vision: Reproducible Infrastructure, Not One-Off Demonstrations

The broader thesis of this paper is that sustainable AI advantage emerges from reproducible infrastructure over individual demonstrations or pilot programs. The Department's history with AI is littered with successful experiments that never scaled: Maven demonstrated computer vision at operational tempo; GIDE experiments proved sensor-to-decision data flow; Agile Flag showed autonomous logistics at the tactical edge. Each was a genuine achievement. None, on its own, produced enduring, enterprise-wide capability because the underlying infrastructure (certification pathways, hardware pipelines, and data architecture) did not exist to replicate and sustain the achievement at scale.

The parallel to defense industrial mobilization is instructive. The Trump administration's use of DPA Title III for critical minerals, the Office of Strategic Capital for supply chain financing, and executive orders linking nuclear energy to AI infrastructure each demonstrate that the authorities exist to build purpose-specific industrial ecosystems when national security demands it. Edge AI requires this same institutional commitment: targeted investment in the specific hardware, certification pathways, and data infrastructure that the tactical edge demands.

The warfighter *needs the model* that is allowed to think, in real time, where it matters most (not necessarily what the latest model is.) Building the infrastructure to deliver this model (repeatedly and at scale) is the defining challenge of defense AI for the next decade. The tools, authorities, and funding mechanisms exist. The task to align them toward edge-first, infrastructure-centric AI architecture (rather than continuing to import commercial capability into environments that structurally reject it) is the challenge that remains.

Appendix 1

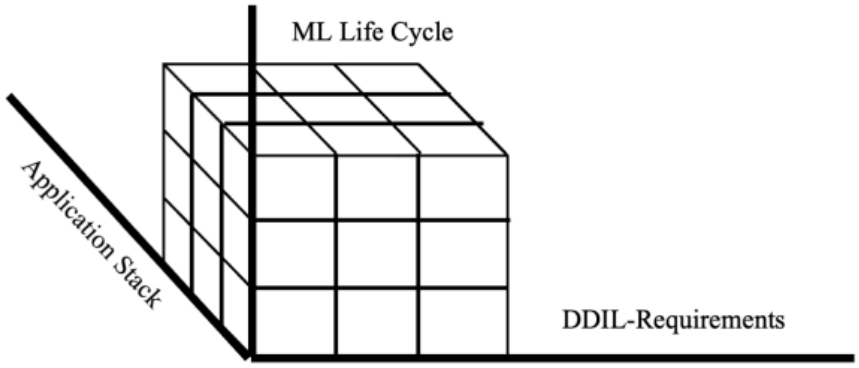


Figure 5. The dimensions of a framework for deploying AI models in DDIL environments

Appendix 2

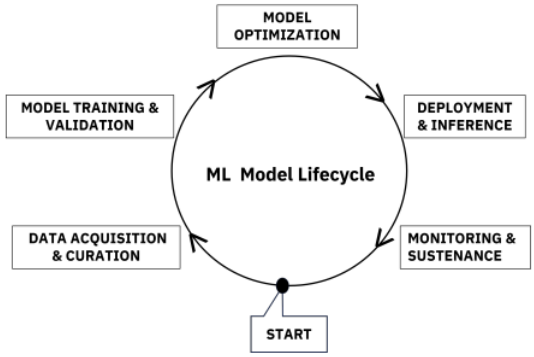


Figure 6. A simplified view of the Machine Learning Model Lifecycle. Various activities are explained in the text below.

Appendix 3

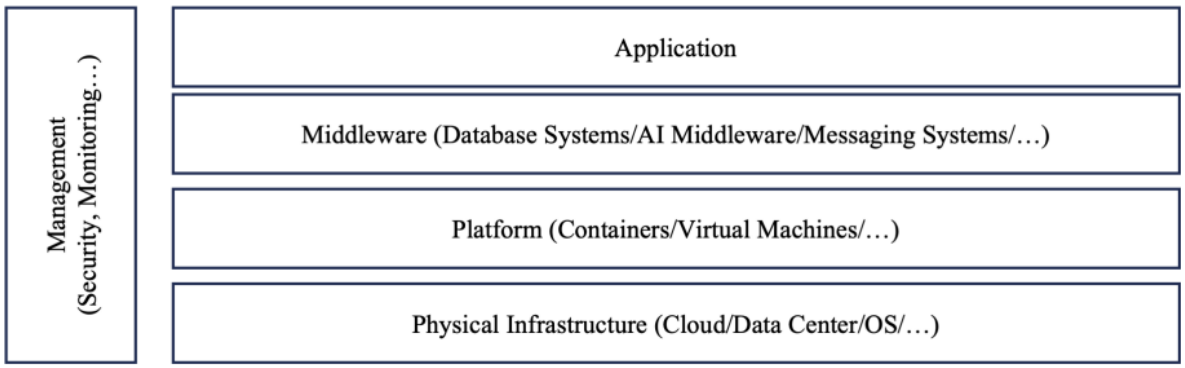


Figure 8. A typical application stack

Works Consulted

1. Rafiq Choudhury, "Project Maven: The Epicenter of US' AI Military Efforts," Interesting Engineering, March 2, 2024, accessed February 25, 2026, https://interestingengineering.com/military/project-maven-the-epicenter-of-us-ai-military-efforts.
2. Department of the Air Force. (2025, April 8). *AFDN 25-1, Artificial Intelligence*. Air Force Doctrine. Retrieved February 25, 2026, from https://www.doctrine.af.mil/Portals/61/documents/AFDN_25-1/AFDN%2025-1%20Artificial%20Intelligence.pdf
3. DoD CDAO. (n.d.). *Initiatives > CJADC2*. Chief Digital and Artificial Intelligence Office. Retrieved February 25, 2026, from https://www.ai.mil/Initiatives/CJADC2/
4. GAO. (2025, April 8). *Defense Command and Control: Further Progress Hinges on Establishing a Comprehensive Framework*. U.S. Government Accountability Office. https://www.gao.gov/products/gao-25-106454
5. Second Front Systems. (2023, April 28). *DoD Impact Level 6 (IL6): What you need to know.* Second Front Systems. Retrieved February 25, 2026, from https://www.secondfront.com/resources/blog/what-you-need-to-know-about-dod-impact-level-6-il6/
6. Palmer, L. (2025, May 28). *Google Distributed Cloud (GDC) & GDC air-gapped appliance achieve DoD Impact Level 6 (IL6) authorization*. Google Cloud Blog.

https://cloud.google.com/blog/topics/public-sector/google-distributed-cloud-gdc-gdc-air-gapped-appliance-achieve-dod-impact-level-6-il6-authorization

7. Air Force Research Laboratory. (2024, September 6). *Autonomous aviation transforms logistics during Agile Flag 24-3 exercise*. https://www.afrl.af.mil/News/Article-Display/Article/3874337/autonomous-aviation-transforms-logistics-during-agile-flag-24-3-exercise/
8. L3Harris Technologies. (n.d.). *High-performance integrated core processor (ICP)*. https://www.l3harris.com/all-capabilities/high-performance-integrated-core-processor-icp
9. U.S. Government Accountability Office. (2023). *Defense logistics: Actions needed to address supply chain risks* (GAO-23-105556). https://www.gao.gov/assets/gao-23-105556.pdf
10. Allen, G. C. (Host). (2024, March 26). *Scaling AI-enabled capabilities at the DOD: Government and industry perspectives* [Event summary]. Center for Strategic and International Studies. https://www.csis.org/analysis/scaling-ai-enabled-capabilities-dod-government-and-industry-perspectives
11. Air Force Research Laboratory. (2013, December 22). *Defense Production Act Title III project results in improved technology for radar and electronic warfare systems* [Press release]. U.S. Department of the Air Force.
12. Choudhury, S. R. (2024, June 10). Project Maven and the future of military AI. *The Washington Post*.
13. Defense Production Act Title III Program. (n.d.). *Defense Production Act Title III*. U.S. Department of Defense.
14. Google Cloud. (2025, December 9). *Chief Digital and Artificial Intelligence Office selects Google Cloud's AI to power GenAI.mil* [Press release].
15. Government Accountability Office. (2025). *Joint all-domain command and control: DOD needs to address data classification challenges to improve command and control interoperability* (GAO-25-106437). U.S. Government Accountability Office.
16. Kazmierczak, T., Akidau, T., Beaumont, C., & Traub, M. (2024). *Continuous delivery for machine learning systems with MLOps* [White paper]. Google Cloud.
17. L3Harris Technologies. (n.d.). *F-35 integrated core processor* [Product datasheet].

18. Second Front Systems. (2023). *The real cost of ATO: Time, money, and mission impact* [White paper].
19. United States Department of Defense, Chief Digital and Artificial Intelligence Office. (2023). *Data, analytics, and artificial intelligence adoption strategy*.
20. United States Department of the Air Force. (2025). *Department of the Air Force data, analytics, and AI for readiness and maintenance* [Fact sheet].
21. Verma, D. C., & Santhanam, P. (2024). MLOps at the edge in DDIL environments. In P. J. Schwartz, B. Jensen, & M. E. Hohil (Eds.), *Artificial intelligence and machine learning for multi-domain operations applications VI* (Vol. 13051, Article 130510V). SPIE. https://doi.org/10.1117/12.3013300
22. White House. (2025, March 20). *Immediate measures to increase American mineral production* [Executive order].
23. White House. (2025, May 23). *Deploying advanced nuclear reactor technologies for national security* [Fact sheet].
24. White House. (2026, January 6). *Prioritizing the warfighter in defense contracting* [Executive order].
25. Dave, R. (2023). *Implementing MLOps on edge-cloud systems: A new paradigm for training at the edge* (Master's thesis). University of Waterloo. https://uwspace.uwaterloo.ca/handle/10012/19709
26. U.S. Department of War. (2023, June 14). *DoD Chief Digital and Artificial Intelligence Office hosts sixth Global Information Dominance Experiment* [Press release]. https://www.war.gov/News/Releases/Release/Article/3427654/dod-chief-digital-and-artificial-intelligence-office-hosts-sixth-global-informa/
27. Caggiano, M. D., Jr. (2024, February 13). *Air Combat Command wraps up Agile Flag 24-1* [News release]. *Air Combat Command, U.S. Air Force*. https://www.acc.af.mil/News/Article-Display/Article/3675401/air-combat-command-wraps-up-agile-flag-24-1/

Endnotes

[i] Department of the Air Force. (2025, April 8). *AFDN 25-1, Artificial Intelligence*. Air Force Doctrine. Retrieved February 25, 2026, from https://www.doctrine.af.mil/Portals/61/documents/AFDN_25-1/AFDN%2025-1%20Artificial%20Intelligence.pdf
[2] Rafiq Choudhury, "Project Maven: The Epicenter of US' AI Military Efforts," Interesting Engineering, March 2, 2024, accessed February 25, 2026, https://interestingengineering.com/military/project-maven-the-epicenter-of-us-ai-military-efforts.
[iii] DoD CDAO. (n.d.). *Initiatives > CJADC2*. Chief Digital and Artificial Intelligence Office. Retrieved February 25, 2026, from https://www.ai.mil/Initiatives/CJADC2/
[iv] Second Front Systems. (2023, April 28). *DoD Impact Level 6 (IL6): What you need to know*. Second Front Systems. Retrieved February 25, 2026, from https://www.secondfront.com/resources/blog/what-you-need-to-know-about-dod-impact-level-6-il6/
[v] Department of Defense. (2022). *DoDI 8510.01: Risk Management Framework (RMF) for DoD IT*. https://www.esd.whs.mil/Portals/54/Documents/DD/issuances/dodi/851001p.pdf
[vi] U.S. Department of War. (2025, December). *The War Department unleashes AI on new GenAI.mil platform* [Press release]. https://www.war.gov/News/Releases/Release/Article/4354916/the-war-department-unleashes-ai-on-new-genaimil-platform/
[vii] Palmer, L. (2025, May 28). *Google Distributed Cloud (GDC) & GDC air-gapped appliance achieve DoD Impact Level 6 (IL6) authorization*. Google Cloud Blog. https://cloud.google.com/blog/topics/public-sector/google-distributed-cloud-gdc-gdc-air-gapped-appliance-achieve-dod-impact-level-6-il6-authorization
[viii] Air Force Research Laboratory. (2024, September 6). *Autonomous aviation transforms logistics during Agile Flag 24-3 exercise*. https://www.afrl.af.mil/News/Article-Display/Article/3874337/autonomous-aviation-transforms-logistics-during-agile-flag-24-3-exercise/
[ix] L3Harris Technologies. (n.d.). *High-performance integrated core processor (ICP)*. https://www.l3harris.com/all-capabilities/high-performance-integrated-core-processor-icp
[x] U.S. Government Accountability Office. (2023). *Defense logistics: Actions needed to address supply chain risks* (GAO-23-105556). https://www.gao.gov/assets/gao-23-105556.pdf
[xi] Allen, G. C. (Host). (2024, March 26). *Scaling AI-enabled capabilities at the DOD: Government and industry perspectives* [Event summary]. Center for Strategic and International Studies. https://www.csis.org/analysis/scaling-ai-enabled-capabilities-dod-government-and-industry-perspectives
[xii] Rafiq Choudhury, "Project Maven: The Epicenter of US' AI Military Efforts," Interesting Engineering, March 2, 2024, accessed February 25, 2026, https://interestingengineering.com/military/project-maven-the-epicenter-of-us-ai-military-efforts.
[xiii] U.S. Department of War. Artificial Intelligence Strategy for the Department of War. Memorandum, January 9, 2026. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF
[xiv] Crystal Group. "The Future of Defense: AI at the Edge for Real-Time Tactical Advantage." April 8, 2025. Accessed March 27, 2026. https://www.crystalrugged.com/the-future-of-defense-ai-at-the-edge-for-real-time-tactical-advantage/ .
[xv] Brandi Vincent, "CDAO Leaves Edge Data Mesh Nodes Behind with Indo-PACOM after Success in Major Exercise," DefenseScoop, May 13, 2025, accessed March 27, 2026, https://defensescoop.com/2025/05/14/cdao-leaves-edge-data-mesh-nodes-indo-pacom-after-major-exercise/
[xvi] Department of the Air Force, Chief Information Officer. (2023). *DAF API reference architecture 2.0*. https://www.dafcio.af.mil/Portals/64/Documents/Strategy/DAF%20API%20Reference%20Architecture%202.0.pdf
[xvii] Aquia Inc. "Breaking the Federal ATO Bottleneck: How Documentation Automation Is Transforming Government Security." September 7, 2025. Accessed March 27, 2026. https://www.aquia.us/blog/breaking-the-federal-ato-bottleneck-how-documentation-automation-is-transforming-government-security.
[xviii] U.S. Department of Defense, Chief Information Officer. Software Modernization Implementation Plan, Fiscal Years 2025–2026. Washington, D.C.: U.S. Department of Defense, April 29, 2025. https://dodcio.defense.gov/Portals/0/Documents/Library/SW-Mod-I-Plan25-26.pdf
[xix] U.S. Department of Defense, Chief Information Officer. Software Modernization Implementation Plan, Fiscal Years 2025–2026. Washington, D.C.: U.S. Department of Defense, April 29, 2025. https://dodcio.defense.gov/Portals/0/Documents/Library/SW-Mod-I-Plan25-26.pdf
[xx] Patrick Tucker, "For DOD, the Future of Large Language Models Is Smaller," Defense One, May 21, 2025, accessed March 27, 2026, https://www.defenseone.com/technology/2025/05/dod-future-large-language-models-smaller/405539/
[xxi] U.S. Department of Defense, Chief Digital and Artificial Intelligence Office. Human Systems Integration Test and Evaluation of Artificial Intelligence-Enabled Capabilities (AIEC): A Framework for the Department of Defense. April 2024. https://www.ai.mil/Portals/137/Documents/Resources%20Page/CDAO_TE_Framework_-_HSI_TES_2024-04-compressed.pdf.
[xxii] NATO. "NATO's DIANA Connects Innovators and End Users through the Rapid Adoption Service." December 11, 2025. Accessed March 27, 2026. https://www.nato.int/en/news-and-events/articles/news/2025/12/12/nato-s-diana-connects-innovators-and-end-users-through-the-rapid
[xxiii] DoD, *"Military and Security Developments Involving the People's Republic of China 2024"* (December 2024 — the annual China Military Power Report). https://defensescoop.com/2025/12/26/dod-report-china-military-and-security-developments-prc-ai-llm/
[xxiv] Recorded Future / Insikt Group, *"Artificial Eyes: Generative AI and China's Military Intelligence"* (2025). https://www.recordedfuture.com/research/artificial-eyes-generative-ai-chinas-military-intelligence
[xxv] Defense One, *"New Products Show China's Quest to Automate Battle"* (March 2025). https://www.defenseone.com/threats/2025/03/new-products-show-chinas-quest-automate-battle/403387/
[xxvi] CSET Georgetown, *"China's Military AI Wish List"* and *"Pulling Back the Curtain on China's Military-Civil Fusion"* (September 2025). https://cset.georgetown.edu/publication/pulling-back-the-curtain-on-chinas-military-civil-fusion/
[xxvii] *"The Private Firms Powering China's Military AI Push"* (The Diplomat, March 2026). https://thediplomat.com/2026/03/the-private-firms-powering-chinas-military-ai-push
[xxviii] Kazmierczak, T., Akidau, T., Beaumont, C., & Traub, M. (2024). *Continuous delivery for machine learning systems with MLOps* [White paper]. Google Cloud. https://services.google.com/fh/files/misc/practitioners_guide_to_mlops_whitepaper.pdf
[xxix] Amazon Web Services. (n.d.). *MLOps foundation roadmap for enterprises*. In *Machine learning best practices for public sector organizations*. AWS. https://docs.aws.amazon.com/whitepapers/latest/ml-best-practices-public-sector-organizations/mlops.html
[xxx] Dwares, J. (2025, January 15). *DoD cloud impact levels explained: IL2, IL4, IL5 & IL6*. Compass IT Compliance. https://www.compassitc.com/blog/understanding-dod-impact-levels-for-cloud-security

[xxxi] Verma, D. C., & Santhanam, P. (2024). MLOps at the edge in DDIL environments. In P. J. Schwartz, B. Jensen, & M. E. Hohil (Eds.), *Artificial intelligence and machine learning for multi-domain operations applications VI* (Vol. 13051, Article 130510V). SPIE. https://doi.org/10.1117/12.3013300

[xxxii] U.S. Department of War. (2023, June 14). *DoD Chief Digital and Artificial Intelligence Office hosts sixth Global Information Dominance Experiment* [Press release]. https://www.war.gov/News/Releases/Release/Article/3427654/dod-chief-digital-and-artificial-intelligence-office-hosts-sixth-global-informa/

[xxxiii] Caggiano, M. D., Jr. (2024, February 13). *Air Combat Command wraps up Agile Flag 24-1* [News release]. *Air Combat Command, U.S. Air Force*. https://www.acc.af.mil/News/Article-Display/Article/3675401/air-combat-command-wraps-up-agile-flag-24-1/

[xxxiv] Aquia Inc. cATO+ and Federal Modernization: Accelerating Risk Management Framework (RMF) Authorizations with Automation. White paper, 2025. https://23570506.fs1.hubspotusercontent-na1.net/hubfs/23570506/Aquia%20cATO+%20and%20Federal%20Modernization%20White%20Paper.pdf.

[xxxv] L3Harris Technologies. "High-Performance Integrated Core Processor (ICP)." Accessed March 27, 2026. https://www.l3harris.com/all-capabilities/high-performance-integrated-core-processor-icp.

[xxxvi] Patrick Tucker, "For DOD, the Future of Large Language Models Is Smaller," *Defense One*, May 21, 2025, accessed March 27, 2026, https://www.defenseone.com/technology/2025/05/dod-future-large-language-models-smaller/405539.

[xxxvii] U.S. Air Force Research Laboratory. "Autonomous Aviation Transforms Logistics during AGILE FLAG 24-3 Exercise." August 23, 2024. Accessed March 27, 2026. https://www.afrl.af.mil/News/Article-Display/Article/3874337/autonomous-aviation-transforms-logistics-during-agile-flag-24-3-exercise/

[xxxviii] Brandi Vincent, "CDAO Leaves Edge Data Mesh Nodes Behind with Indo-PACOM after Success in Major Exercise," DefenseScoop, May 13, 2025, accessed March 27, 2026, https://defensescoop.com/2025/05/14/cdao-leaves-edge-data-mesh-nodes-indo-pacom-after-major-exercise/.

[xxxix] Dinesh C. Verma and P. Santhanam, "MLOps at the Edge in DDIL Environments," IBM Research, 2024, https://research.ibm.com/publications/mlops-at-the-edge-in-ddil-environments.

[xl] U.S. Department of War. *Artificial Intelligence Strategy for the Department of War*. Memorandum, January 9, 2026. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF.

[xli] Patrick Tucker, "New Pentagon Program to Speed Up Software Acquisition Set to Launch," *DefenseScoop*, April 28, 2025, accessed March 27, 2026, https://defensescoop.com/2025/04/29/dod-cio-katie-arrington-swift-software-acquisition-ato.

[xlii] White House. "Immediate Measures to Increase American Mineral Production." Executive Order, March 20, 2025. https://www.whitehouse.gov/presidential-actions/2025/03/immediate-measures-to-increase-american-mineral-production/

[xliii] White House. "Restoring America's Maritime Dominance." Executive Order, April 9, 2025. https://www.whitehouse.gov/presidential-actions/2025/04/restoring-americas-maritime-dominance/

[xliv] Air Force Research Laboratory. "Defense Production Act Title III Project Results in Improved Technology for X-Band GaN MMICs." December 23, 2013. Accessed March 27, 2026. https://www.wpafb.af.mil/News/Article-Display/Article/819308/defense-production-act-title-iii-project-results-in-improved-technology-for-x-band-gan-m/

[xlv] U.S. Department of Defense, Office of Strategic Capital. "Office of Strategic Capital Announces First Notice of Funding Availability to Secure the U.S. Industrial Base." September 30, 2024. Accessed March 27, 2026. https://www.war.gov/News/Releases/Release/Article/3921005/office-of-strategic-capital-announces-first-notice-of-funding-availability-to-secure-the-us-industrial-basese/

[xlvi] U.S. Department of War. *Acquisition Transformation Strategy: Rebuilding the Arsenal of Freedom*. November 10, 2025. https://media.defense.gov/2025/Nov/10/2003819441/-1/-1/1/ACQUISITION-TRANSFORMATION-STRATEGY.PDF

[xlvii] K&L Gates LLP. "Key Provisions on Artificial Intelligence in Fiscal Year 2025 National Defense Authorization Act." January 1, 2025. Accessed March 27, 2026. https://www.klgates.com/Key-Provisions-on-Artificial-Intelligence-in-Fiscal-Year-2025-NDAA-1-2-2025

[xlviii] U.S. Department of War. *Artificial Intelligence Strategy for the Department of War*. Memorandum, January 9, 2026. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/ARTIFICIAL-INTELLIGENCE-STRATEGY-FOR-THE-DEPARTMENT-OF-WAR.PDF

Detecting Systematic Infrastructure Attacks via Geospatial Intelligence

Kevin Chen and Cole Griffiths

Jackson School of Global Affairs, Yale University

Schmidt Program on AI, Emerging Technologies, and National Power

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Bios

Kevin Chen is a Master of Public Policy candidate at the Yale Jackson School of Global Affairs, focusing on AI influence operations, technology governance, and Indo-Pacific geopolitics. Prior to Yale, he previously served in the U.S. intelligence community as a data scientist. A graduate of Dartmouth College, Kevin is fluent in Mandarin Chinese and seeks to advance democratic resilience through emerging technology policy and defense innovation.

Cole Griffiths is a second year MPP at the Yale Jackson School of Global Affairs. In his career, Cole leverages emerging technology to promote human security. He has worked in policy and research roles with Stanford CISAC, the Hoover Institution, the UN DPPA Innovation Cell, and the US State Department. A graduate of Stanford University, Cole concentrates on issues related to geospatial technology, AI trust and safety, and conflict risk analysis.

Abstract

Modern warfare increasingly targets infrastructure indispensable to civilian survival, particularly water and food systems, as a tactical instrument of coercion. Despite extensive reporting, a capability gap persists in scalable tools for real-time damage detection and humanitarian impact prediction. This paper introduces an integrated geospatial artificial intelligence (GeoAI) framework leveraging multi-modal satellite imagery, computer vision, and data fusion to map vulnerabilities in infrastructure and supply chains. By fusing pixel-level change detection with ACLED conflict data and WFP logistics telemetry, we generate spatially explicit predictive risk indicators. We evaluate this toolkit through case studies on the destruction of the water-energy nexus in Ukraine and the food system blockades in Yemen. Contributions include a scalable OSINT toolkit, a multi-tiered early-warning typology, and a prototype dashboard architecture for crisis management. Finally, we address ethical considerations regarding demographically identifiable information (DII), demonstrating how bridging GeoAI with humanitarian analysis advances early-warning capabilities for national security and global response.

Keywords: Geospatial Intelligence, Infrastructure Monitoring, Conflict Zones, Humanitarian Assistance, Early-Warning Systems, Remote Sensing, Computer Vision

1 Introduction: Humanitarian Innovation in Modern Warfare

The character of modern warfare is undergoing a profound transformation, increasingly defined by the systematic and deliberate targeting of civilian infrastructure. In contemporary conflicts, the traditional battlefield expands into the urban and logistical domains, fundamentally altering the strategic landscape of national and global security. State and non-state actors alike weaponize access to foundational human needs, targeting water treatment facilities, dams, agricultural storage sites, and humanitarian logistics networks. Degradation of objects indispensable to civilian survival serves as an instrument of coercion, demographic engineering, and hybrid conflict, amplifying civilian harm and accelerating the onset of complex humanitarian emergencies. Recent theaters of conflict demonstrate a disturbing operational pattern in which the destruction of critical infrastructure is not merely collateral damage, but a primary tactical objective designed to break societal resilience, exhaust state resources, and force political capitulation.

While international humanitarian law establishes unequivocal prohibitions against such actions, the ”fog of war” generally hinders legal enforcement and the execution of timely humanitarian interventions. Policymakers, intelligence agencies, and humanitarian organizations consistently face a severe information deficit, lacking scalable, systematic tools to detect infrastructure damage and forecast downstream humanitarian impacts in near-real time. Traditional ground-based assessments are frequently rendered obsolete, delayed, or entirely perilous by active hostilities, access constraints, improvised explosive devices (IEDs), and the sheer scale of destruction across contested territories. Consequently, there is an urgent operational requirement for advanced technological solutions capable of bridging the gap between strategic early warning and tactical humanitarian response.

This manuscript presents an exhaustive analysis of an integrated Geospatial Artificial Intelligence (GeoAI) framework designed to address this critical capability gap. Moving beyond retrospective damage detection, the architecture introduces a typology of early-warning indicators, transforming raw remote sensing data and unstructured conflict event logs into spatially explicit, predictive risk indices. Through case studies of the water infrastructure crisis in Ukraine and the food supply chain degradation in Yemen, this analysis demonstrates the operational efficacy of the GeoAI toolkit. Ultimately, this research underscores how the responsible application of emerging defense technologies, specifically artificial intelligence, multi-modal remote sensing, and data

fusion, can enhance global security resilience and mitigate destabilizing threats to human survival on the front lines of modern conflict.

2 The Historical Evolution of Earth Observation in Conflict Monitoring

Contextualizing the historical evolution can help better understand the significance of applying GeoAI to civilian infrastructure protection. For decades, satellite Earth observation was the exclusive domain of state military and intelligence apparatuses, heavily restricted by security classifications and exorbitant costs [1]. During the Cold War, the primary application of satellite imagery was strategic reconnaissance: monitoring troop mobilizations, nuclear silo construction, and naval fleet movements [1].

Beginning in the 1990s, the commercialization and privatization of the space sector initiated a paradigm shift, making high-resolution satellite imagery increasingly accessible to non-state actors. Proliferation of commercial satellite constellations by firms like Maxar,
Planet Labs, and Airbus dramatically transformed remote sensing on multiple dimensions. Private satellites now rival or exceed governments' capabilities in spatial resolution (level of ground detail captured per/pixel), global coverage, temporal resolution and revisiting rates, and hyperspectral sensing. Where high-resolution imagery access used to be walled behind export controls and shutter limitations, individual consumers can now easily shop a global marketplace for imagery products. Nonprofits, researchers, and human rights organizations particularly benefit from this widespread availability. Today, mainline organizations like Amnesty International and independent investigative collectives routinely utilize satellite imagery to verify military attacks, investigate alleged war crimes, and track the destruction of civilian infrastructure in regions including Darfur, Myanmar, and Gaza.

The ongoing conflict in Ukraine further pushed the role of commercial satellite imagery into the global spotlight [1]. Prior to the February 2022 invasion, open-source intelligence (OSINT) analysts utilized commercial imagery to track the buildup of Russian forces, preempting official state intelligence releases and shaping international public opinion [1]. This democratization of intelligence represents a critical advancement; however, it also introduces a massive data processing bottleneck. The sheer volume of imagery generated daily by commercial constellations vastly exceeds the analytical capacity of human analysts: estimates suggest Planet offers more than 500 petabytes of satellite data for users [2]. This specific bottleneck necessitates the integration of

Artificial Intelligence and deep learning for automating detection of infrastructure degradation at a planetary scale.

3 Legal and Policy Frameworks: Civilian Protection in the Digital Age

International humanitarian law provides the foundational architecture for protecting civilian infrastructure during armed conflict. The evolving nature of warfare demands a rigorous application of treaty law and customary practice, particularly as emerging technologies complicate the distinction between military objectives and civilian assets.

3.1 Article 54 and Customary International Law

Specifically, Article 54(2) of the 1977 Additional Protocol I to the Geneva Conventions establishes the absolute protection of objects indispensable to the survival of the civilian population [3]. These provisions explicitly prohibit any military action designed to ”attack, destroy, remove or render useless” assets such as foodstuffs, agricultural areas, crops, livestock, drinking water installations, and irrigation works for the specific purpose of denying them for their sustenance value [3]. This principle is universally recognized as a norm of customary international law, applicable in both international and non-international armed conflicts. Military operations involving collateral deprivation are generally deemed unlawful if the primary object is to starve the civilian population or force its displacement [4].

This customary norm is deeply embedded in state practice and military manuals around the world. For example, Canada’s Law of Armed Conflict at the Operational and Tactical Level manual explicitly prohibits attacking, destroying, or rendering useless such survival objects ”whatever the motive” [4]. Despite not ratifying Additional Protocol I, the US accepts Article 54(2) as binding customary law [5]. Similar commitments are not just confined to wealthy Western states or democracies either. Burundi’s Regulations on International Humanitarian Law prohibits targeting drinking water installations, harvests, and livestock for the sole purpose of starving civilians [4]. The People’s Republic of China accepts binding norms against weaponized starvation in warfare [6]. Exceptions exist only if these objects are used solely to sustain combatants or in direct support of military action; even then, treaty law only permits these exceptions if prospective attacks will not leave the civilian population with inadequate food or water, thereby causing starvation [3].

3.2 The Integration of Remote Sensing for Legal Accountability

Integration of GeoAI into conflict monitoring provides a revolutionary mechanism for documenting and analyzing potential violations of these norms. Historically, the legal analysis of environmental and infrastructure protection during warfare focused almost exclusively on post-conflict pollution and direct military damage, often assessed years after hostilities concluded [7]. However, the concurrent application of international human rights law and environmental law during armed conflicts mandates that states must take environmental and civilian survival considerations into account when assessing military necessity and proportionality in real time [7].

The International Law Commission's Draft Principles on the Protection of the Environment in Relation to Armed Conflict reflect this broader framing, shifting the focus beyond traditional international humanitarian law to include the role of natural resources as drivers of conflict and the protection of environments before, during, and after hostilities [7]. Highresolution satellite imagery, coupled with AI-driven change detection, offers an immutable, objective record of infrastructure degradation over time. This capability not only supports immediate humanitarian triage but also facilitates long-term accountability, providing empirical, timestamped evidence of systematic attacks on water and food systems that contravene Article 54.

4 Analytical Framework and Algorithmic Pipelines

To operationalize this capability, our project relies on a multi-tiered technical architecture merging remote sensing, deep learning, and open-source data fusion. The objective is to automate the extraction of meaningful intelligence from vast streams of unstructured data, achieving a high degree of precision in environments characterized by noise, obfuscation, and rapid geographical change.

4.1 Multi-Modal Earth Observation Acquisition

The efficacy of any geospatial artificial intelligence framework is fundamentally contingent upon the quality, resolution, and temporal frequency of its foundational data layers. Singlesensor approaches are inherently flawed in conflict zones; optical sensors are blinded by cloud cover or

deliberate smoke screens, while radar sensors struggle with complex urban geometries. Therefore, the system ingests data from a constellation of Earth Observation platforms, utilizing diverse spectral modalities.

Optical imagery forms the baseline for computer vision models, providing the detailed structural context required for identifying specific ground object types. To ensure accurate detection of structural damage caused by shelling, it is essential to capture both top-down and oblique (30-50 degrees from axis) very-high resolution imagery, avoiding the omission of buildings that have intact roofs but destroyed load-bearing walls [8].

SAR is particularly critical for infrastructure monitoring due to its structural sensitivity. Partial building collapse or the destruction of a pipeline changes the position of the radar scatterer within an image region. An optical sensor looking directly down (or nadir) might miss this. But SAR images allow detection detection of changes to the structure through phase information (or coherence). Furthermore, SAR excels at detecting localized flooding caused by damaged dams, as water surfaces act as specular reflectors, yielding distinct dark backscatter signatures that algorithms can isolate regardless of weather [9].

Nighttime Light imagery serves as an accurate proxy for power grid functionality. Comparing conflict-time NTL images against normal baseline conditions allows for the rapid generation of high-resolution outage maps, which in turn indicate the likely failure of energydependent water pumping stations [10].

4.2 Deep Learning Architectures and Semantic Segmentation

The analytical core of the framework employs advanced deep learning architectures to process ingested imagery. Traditional remote sensing methods based on rule-based algorithms, thresholding spectral indices, or manual feature extraction struggle with poor generalization, high computational requirements, and an inability to adapt to complex urban environments with varying spectral reflectance [11].

To address these limitations, the framework utilizes Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs) configured for semantic segmentation and object detection . Models such as U-Net, ResNet, SegFormer, and YOLOv11 assign categorical labels to pixels or generate highly accurate bounding boxes around damaged structures [11].

In tasks involving the detection of armed conflict damages in medium-resolution satellite

Sensor Modality	Primary Characteristics	Application in Infrastructure Monitoring	Constraints
Very-High-Resolution (VHR) Optical	Sub-meter spatial resolution (e.g., Maxar, Airbus); captures visible and near-infrared (NIR) spectra.	Granular damage assessment of buildings, pumping stations, and greenhouses.	Heavily limited by cloud cover, atmospheric haze, and daylight dependency.
Synthetic Aperture Radar (SAR)	Active microwave sensing (e.g., Sentinel-1); penetrates clouds and operates day or night.	Detecting flooded areas from dam breaches; mapping structural changes via phase coherence and backscatter anomalies.	Highly complex data processing; speckle noise can obscure fine geometric details.
Nighttime Light (NTL) Radiometry	Detects artificial illumination emissions; 30m to 500m resolution.	Proxy for power grid functionality, urban displacement, and blackout mapping.	Susceptible to lunar illumination interference, snow reflection, and viewing angle distortion.

imagery (e.g., Sentinel-2), maintaining spatial resolution throughout the network is critical [12]. Damaged objects often register as only a few pixels or even sub-pixel in size. Empirical studies demonstrate that preserving striding operations at a value of 1 prevents the loss of crucial spatial features, significantly outperforming models with higher abstraction and larger strides (such as standard stride-2 U-Nets) [12]. To counteract class imbalance, where damaged pixels are vastly outnumbered by non-damaged terrain, advanced implementations weight the loss function during training and artificially expand ground truth labels (e.g., buffering a single damage coordinate to a nine-pixel area) to favor the avoidance of false negatives [12].

Advanced object detection models like YOLOv11 demonstrate exceptional capability for real-time infrastructure assessment. Research utilizing enhanced YOLO architectures (such as YOLO-

ATL) incorporating Global Attention Mechanisms and multi-scale feature fusion has shown significant improvements in detecting small, complex damages to infrastructure like roads and bridges, achieving high accuracy while maintaining rapid inference speeds suitable for field deployment.

4.2.1 Mathematical Foundations and Evaluation Metrics

Our framework rigorously evaluates the performance of these deep learning models using standard statistical metrics derived from the confusion matrix: True Positives (TP), False Positives (FP), True Negatives (TN), and False Negatives (FN). In the context of conflict monitoring, prioritizing high Recall is often more critical than absolute Precision. Missing a damaged water facility (a False Negative) carries a higher humanitarian cost than manually verifying a false alarm (a False Positive). The primary metrics used to benchmark model
efficacy are:

Precision: The ratio of correctly predicted positive observations to the total predicted positives. It indicates the accuracy of the damage alerts.

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

Recall (Sensitivity): The ratio of correctly predicted positive observations to all observations in the actual class. It measures the model's ability to find all damaged infrastructure.

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

F1-Score: The weighted average of Precision and Recall. Therefore, this score takes both false positives and false negatives into account, making it highly useful for datasets with severe class imbalances (typical in conflict zones).

$$F1 = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

Intersection over Union (IoU): Also known as the Jaccard Index, this is a critical metric

for semantic segmentation, measuring the overlap between the predicted damage mask and the ground truth mask.

$$IoU = \frac{TP}{TP + FP + FN} \tag{4}$$

4.2.2 Foundation Models and Zero-Shot Generalization

Recent advancements emphasize the integration of large-scale Geospatial Foundation Models (GFMs). The Segment Anything Model (SAM) in particular, provides robust zero-shot generalization capabilities [13]. When adapted for geospatial applications via frameworks like SAMUnet or segment-geospatial (samgeo), these models can extract high-fidelity masks of flooded areas or damaged infrastructure without requiring extensive, manually annotated training datasets for every new conflict theater [13]. Furthermore, multimodal GFMs like ESA-IBM's TerraMind demonstrate high recall when fine-tuned on harmonized optical and SAR datasets for global-scale, near-real-time disaster mapping [9]. Unsupervised anomaly detection using Variational Auto-Encoders (VAEs) applied to high-cadence commercial imagery from providers like Planet also allows for the near-real-time detection of conflict-related fires and burn scars without relying on extensive ground-truth labels.

5 Open-Source Intelligence and Geospatial Data Fusion

Pixel-level analysis, while technologically sophisticated, is insufficient for comprehensive strategic intelligence if decoupled from socio-political ground truth. The framework achieves true geospatial fusion by layering algorithmic remote sensing outputs over dynamic humanitarian, logistical, and conflict datasets [14].

5.1 The Armed Conflict Location & Event Data Project (ACLED)

Our primary vector for conflict contextualization is the Armed Conflict Location & Event Data Project (ACLED). ACLED provides highly granular, real-time records of global political violence

and demonstrations [15]. ACLED data is disaggregated down to the exact day, latitude, and longitude, capturing specific violent event types and the distinct actors involved [15].

To operationalize this raw data within a geographic information system, our framework utilizes techniques like Kernel Density Estimation (KDE) to transform discrete, vector-based conflict incident points into continuous raster heatmaps [16]. By setting a search radius and utilizing fatalities as a weight field, the assigned pixel values correspond to the intensity of conflict, clearly demarcating the borders of high-risk zones [16].

We further leverage ACLED data for predictive modeling through the ACLED Conflict Alert System (CAST) [17]. CAST employs a sophisticated machine learning pipeline, testing hundreds of Random Forest and XGBoost models to forecast political violence at the subnational (first administrative division) level for a rolling six-month period [17]. These models are trained using a combination of ”simple” predictors (recent violence trends) and ”complex” predictors, which incorporate actor interactions, strategic developments (e.g., peace agreements), and demographic proxies from sources like WorldPop [17]). Since training data are partitioned using time-series cross-validation splits, the models are prevented from overemphasizing short-term anomalies [17]. By generating discrete forecasts of violent events rather than mere probabilities, CAST provides actionable foresight [17]. When these predictive conflict zones geographically intersect with critical water or food infrastructure nodes detected via satellite imagery, the framework proactively flags the area for impending humanitarian risk [18]. Research confirms that integrating quantitative conflict event data with machine learning significantly improves the accuracy of models predicting food security outcomes [18].

5.2 Humanitarian Logistics and Supply Chain Telemetry

Our framework also ingests logistical and supply chain data from the World Food Programme (WFP) and associated humanitarian clusters [19]. This intelligence stream tracks the operational status of international dry ports, the movement of humanitarian convoys, warehouse storage capacities, and market price anomalies [19].

For example, the WFP Logistics Cluster conducts regular Warehouse Capacity Mapping surveys, identifying exact shareable space across active conflict governorates, tracking total warehouse capacity and immediately available shareable space for incoming aid [19]. WFP monitors whether specific entry points, border crossings, or maritime ports are open, closed, or temporarily suspended due to geopolitical tension [19]. The fusion of these multi-dimensional data

layers, geographic movement, temporal conflict patterns, and organizational networks, reveals previously obfuscated ”gray zone” activities [14]. Analysts can triangulate ACLED conflict data with changes in WFP convoy routes and real-time remote sensing to identify how kinetic events precipitate acute logistical bottlenecks, allowing them to anticipate crises before supplies run out.

6 A Typology of Early-Warning Indicators for Civilian Harm

To transition from retrospective damage assessment to anticipatory intelligence, the framework codifies a specific typology of early-warning indicators. Drawing upon principles of conflict forecasting, disaster risk reduction, and food security analysis, these indicators are structured into three distinct analytical horizons: Structural, Proximate, and Trigger factors [20].

The application of this typology is highly modular and data-driven. Historical validation in conflict settings demonstrates quantitative warnings can be reliably issued before a sharp deterioration in food security occurs [22]. With optimally-set thresholds capturing inflation,

Indicator Classification	**Temporal Horizon & Definition**	**Geospatial / OSINT Data Signatures**	**Downstream Humanitarian Impact**
Structural	Long-term: Underlying vulnerabilities and systemic conditions making populations susceptible to crises [20].	Chronic water scarcity indices; historical reliance on single-point infrastructure; sustained agricultural stress (low NDVI over multiple seasons) [21].	Creates a baseline fragility; populations possess diminished coping mechanisms against subsequent economic or kinetic shocks.

Proximate	Medium-term: Accelerators that elevate risk, often observable months prior to acute crises ([20]).	Widespread power grid degradation (NTL reduction); gradual depletion of strategic grain reserves; macroeconomic indicators (currency devaluation, Minimum Food Basket cost spikes) [21].	Degrades the capacity of utility providers and humanitarian actors to sustain baseline service delivery; prompts negative coping strategies [21].
Trigger	Short-term: Catalyst events that precipitate immediate humanitarian emergencies and displacement [20].	Kinetic strikes on pumping stations (AI damage detection); immediate flooding from ruptured dams (SAR); sudden port closures or convoy blockades (ACLED/WFP data) [9]; [19].	Acute denial of objects indispensable to survival; triggers rapid displacement, extreme malnutrition, and waterborne disease outbreaks.

conflict proximity, and agricultural productivity shocks, indicators can signal deterioration most accurately at a five-month lead time, greatly enhancing the operational window for anticipatory humanitarian action [22]. Crucially, mitigating the blind spots of purely remote data requires effective early warning systems to integrate civilian-led inputs and local knowledge, ensuring that the indicators are culturally relevant and responsive to ground realities
[23].

7 Case Study I: Ukraine – Water Infrastructure and the Energy Nexus

The armed conflict in Ukraine serves as a primary case study for the systematic degradation of civilian infrastructure and the absolute necessity of geospatial artificial intelligence for near-real-time monitoring. In 2024 and 2025, the harm to the civilian population worsened demonstrably due to sustained and systematic attacks on Ukraine's energy infrastructure
[24].

7.1 Systemic Destruction and the Humanitarian Toll

The Ukrainian scenario highlights the profound interdependencies between energy and water infrastructure. By the end of 2024, the total damage to Ukrainian energy infrastructure was estimated at $14.6 billion, representing a massive 46% dynamic increase from earlier assessments [25]. Massive attacks disabled over 4 gigawatts of generation capacity, critically damaging thermal and hydroelectric power plants [25]. This systemic destruction of the energy grid directly incapacitated municipal water supply and sewerage systems, causing a secondary humanitarian crisis. Ukrainian water systems are highly energy-dependent, historically consuming 1.01 kWh/m^3 of energy, nearly double the European average [26].

Consequently, independent experts estimate physical damage to the Ukrainian water sector specifically at $4.6 billion, with approximately 39,700 kilometers of water networks damaged or destroyed [26]. The humanitarian fallout is severe: currently, only 68% of the population maintains access to centralized water supply, leaving 1.7 million children without safe water access [26]. The destruction of the Kakhovka Hydroelectric Power Plant in 2023 stands as a catastrophic Trigger indicator, resulting in billions of dollars in multi-sectoral damages and the immediate denial of drinking water to over a million people [25] [26]. Furthermore, water quality has plummeted, with one-third of all drinking water samples failing to meet national standards, drastically elevating the risk of waterborne diseases [26].

As of early 2025, the total cost of reconstruction and recovery in Ukraine over the next decade is projected to reach an astounding $524 billion [27].

7.2 Algorithmic Damage Assessment and Resilience Planning

In environments where ground access is highly restricted due to active fighting and extensive mine contamination, the deployment of automated AI damage assessment is critical. Commercial entities like Tensorflight have deployed deep learning neural networks across cities like Bucha, Irpin, and Mariupol to process high-resolution optical satellite imagery [8]. Utilizing both top-down and oblique imagery, the AI identifies building footprints, classifies structures along a severity gradient (e.g., intact, suitable for repair, completely demolished), and cross-references historical data to estimate reconstruction costs [8].

Figure 1

Visualization of Kharkiv damage assessment enabled by an xBD model implementation



Within our proposed framework, this pipeline is specifically calibrated to monitor water utility sites and reservoirs. However, GeoAI is equally crucial for planning resilient recovery. A primary example is the emergency restoration of the Mykolaiv drinking water pipeline. After Russian forces destroyed the previous Dnipro River water supply in 2022, the Ukrainian government initiated a massive $210.5 million (UAH 8.7 billion) project to build an entirely new 130+ km pipeline drawing from the Pivdennyi Buh River. Leveraging modern infrastructure design informed by threat modeling, the new system places the bulk of the pipes, water pumping stations, and critical cable networks deep underground to protect them from frequent Russian shelling. Air raid shelters were built for personnel, and controls are sited remotely with reserve power supplies. Completed ahead of schedule in August 2025 and under budget, the system now provides 120,000 cubic meters of fresh water per day to half a million residents at a cost savings of UAH 2.4 billion.

8 Case Study II: Yemen – Monitoring Food Supply Chain Degradation

In contrast to the rapid, kinetic destruction of infrastructure observed in Ukraine, the crisis in Yemen illustrates a protracted, systemic degradation of the food supply chain compounded by severe climate-induced vulnerabilities, economic collapse, and geopolitical blockades [21] [22]. Yemen represents one of the world's most severe protracted humanitarian crises, requiring a GeoAI approach focused heavily on logistical monitoring, agricultural stress detection, and OSINT economic tracking.

8.1 Blockade Dynamics and Protracted Starvation

Yemen is fundamentally water-scarce and historically reliant on imports for approximately 85% to 90% of its staple foods. The conflict has systematically choked the logistical arteries necessary for food distribution. Escalating regional conflict and the shipping crisis in late
2024 through 2025 compounded damage to port infrastructure along the Red Sea. This resulted in a notable decline of 13% in food imports and 28% in fuel imports through Sana'a Based Authorities (SBA) controlled ports between January and November 2025. While the internationally recognized Government of Yemen (IRG) implemented economic measures in August 2025 that allowed the Aden-based Yemeni Rial (YER) to appreciate to roughly 1,616 YER/USD, the cost of the Minimum Food Basket (MFB) remains prohibitively high for most citizens.

By the end of 2025, 61% of surveyed households nationwide struggled to meet their minimum food needs, with 35% facing severe food deprivation. The severity of hunger is disproportionately higher in SBA areas due to suspended assistance and limited livelihoods [21]. The Integrated Food Security Phase Classification (IPC) projected that by early 2026, over 5.38 million people (53% of the analyzed population in GoY areas alone) would face IPC
Phase 3 (Crisis) or above levels of acute food insecurity, including over 1.6 million in Phase 4 (Emergency) [28]. A crippling funding shorfall severely exacerbates this crisis. Incoming contributions to the WFP declined by over 70% from 2024 to 2025, forcing the WFP to reduce the number of targeted emergency food assistance beneficiaries in IRG areas from 3.4 million down to just 1.6 million people [21]. Yemen remains at a humanitarian precipice even as the international capacity to respond is declining.

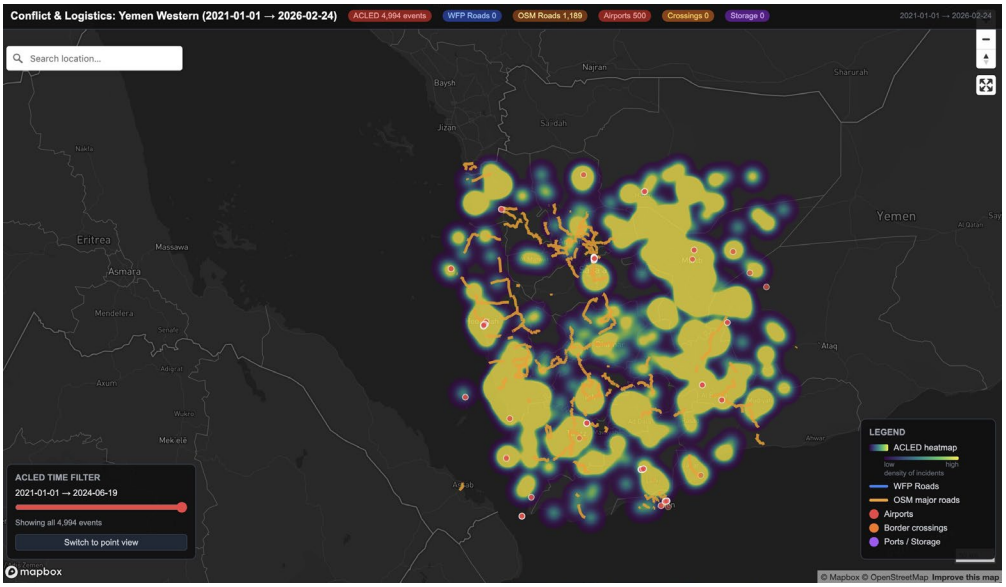
8.2 Integrating Satellite Intelligence with WFP Logistics Data

Monitoring this degradation requires fusing distinct, multi-scalar data streams. The GeoAI framework utilizes multispectral satellite imagery to calculate agricultural stress indicators (like the Normalized Difference Vegetation Index, NDVI) and track climate shocks, such as the delayed and insufficient rainfall in April 2025 that caused late planting and significant crop losses [21] [28] [29].

Simultaneously, our framework ingests data from the WFP Logistics Cluster, which coordinates access constraints, tracks the operational status of international dry ports (e.g., Al-Wadea, Shahen), and monitors available warehouse space [19]. For instance, when shipping companies suspend operations to Al Hodeidah and Aden Ports due to Red Sea security threats, or when internal connecting roads reopen (such as the Al-Dhalea to Sana'a route reopening in May 2025 after a seven-year closure), the Logistics Cluster records these vital access variables [19].. By fusing this logistical ground-truth with ACLED conflict data and AI-derived agricultural assessments, the system generates highly accurate, localized predictions of food security emergencies, allowing humanitarian actors to optimize the routing of shrinking aid resources [22].

Figure 2

Visualization of conflict events dashboard overlaid with WFP logistics corridors.



9 Ethical Considerations and Data Privacy

The deployment of a framework like GeoAI in conflict zones introduces profound ethical considerations that must be rigorously managed. The same satellite imagery and machine learning models used to optimize humanitarian aid delivery can be co-opted by malicious actors to target vulnerable populations.

A primary concern is the protection of Demographically Identifiable Information (DII) [30]. While individual identities may not be visible from space, high-resolution remote sensing can identify the locations, movement patterns, and infrastructure reliance of specific ethnic, religious, or displaced groups. When disparate datasets, such as mobile phone telemetry, social media sentiment, and satellite imagery, are fused, it creates a ”mosaic effect,” unintentionally revealing sensitive knowledge that compromises group privacy and safety [30].

Satellite imagery and digital ground truth also face dangers of spoofing, hallucination, or manipulation from generative AI. Misinformation risks always abound when collating open-source evidence in a world of image GANs. GeoAI is not immune from these manipulation concerns. Researchers have demonstrated the possibility for ”deepfake cartography,” spoofing false satellite imagery that can be credibly ingested into models or GIS [31].

Furthermore, algorithms trained on non-representative data can suffer from biases involving missing or underrepresented categories, leading to the misallocation of life-saving resources [30]. As GeoAI systems scale, humanitarian actors must implement strict ethical guardrails, including differential privacy techniques, data access limitations, and explainable AI (XAI) models, ensuring that these tools serve as a shield for civilian protection rather than a weapon for surveillance and targeting.

10 Design-Oriented Policy: The Humanitarian Dashboard

The sheer volume and complexity of the intelligence generated by this GeoAI framework necessitate a highly optimized user interface. Without effective visualization, machine learning outputs and predictive risk indicators remain siloed within academic or intelligence enclaves, inaccessible to the policymakers, warfighters, and frontline operators who require them most. The final component of the proposed framework is the development of a prototype interactive dashboard designed specifically for national security and humanitarian decision-makers

[32].

10.1 Principles of Geospatial Dashboard Design

To ensure the dashboard functions as an effective instrument for ”Data for Action” [33], its architecture adheres to strict UI/UX design principles tailored for crisis management and cognitive ease:

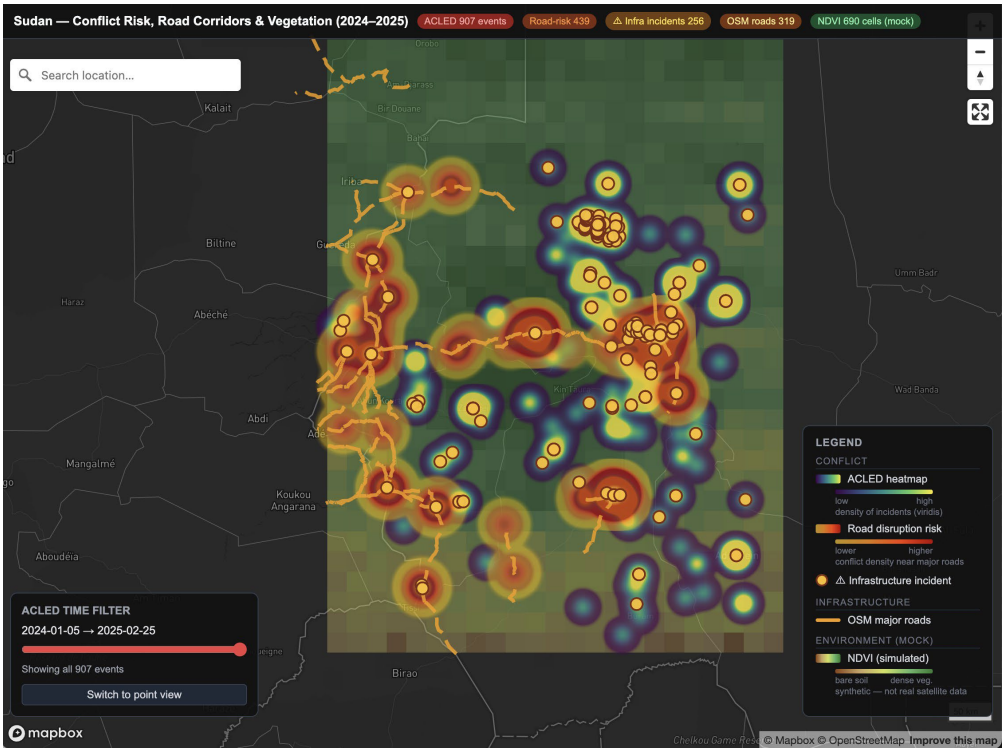
Design Principle	Implementation Strategy	Benefit for Decision-Makers
The 5-Second Rule	Structuring the interface so users can comprehend the most critical insights (e.g., total populations at risk) within five seconds of viewing.	Minimizes cognitive load in high-stress, time-sensitive emergency management environments.
The Inverted Pyramid	Displaying macro-level insights at the top, mid-level trends in the center, and granular, interactive map details at the bottom.	Aligns with natural information consumption patterns, allowing executives to grasp the overview before analysts drill into specifics .
Progressive Disclosure	Presenting a generalized overview initially, with deep-dive data (e.g., semantic segmentation polygons of damaged greenhouses) revealed only upon user interaction.	Prevents visual clutter and overwhelming the user, maintaining a clean interface while preserving data depth.
Coordinated Multiple Views (CMV)	Linking geospatial maps dynamically with non-spatial data charts. Filtering an ACLED timeline automatically updates the map to show concurrent infrastructure damage [34].	Facilitates rapid contextualization and pattern recognition between kinetic events and humanitarian outcomes [34].

By transforming disparate data streams into a cohesive, interactive narrative, the dashboard bridges the gap between technical remote sensing analysis and operational humanitarian deployment. It avoids visual clutter by limiting the interface to a concise set of purpose-driven

visualizations, utilizing drill-downs and filters [32]. Crucially, the dashboard explicitly maps the early-warning typology, utilizing high-contrast visual cues to direct the user's attention immediately to areas requiring urgent resource allocation or diplomatic signaling [32].

Figure 3

Integrative dashboard fusing and conveying multi-stream humanitarian data



11 Conclusion

The intersection of emerging defense technologies, remote sensing, and open-source intelligence presents an unprecedented opportunity to redefine how the international community monitors, understands, and responds to the evolving nature of warfare. As demonstrated by the devastating kinetic impacts on water networks in Ukraine and the systemic collapse of the food supply chain in Yemen, the deliberate targeting of civilian infrastructure represents a defining, and deeply concerning, characteristic of modern conflict.

The GeoAI framework outlined in this report offers a robust, scalable solution to the persistent challenges of data scarcity and operational opacity in active war zones. By synthesizing very-high-resolution optical and radar satellite imagery with advanced computer vision algorithms, ranging from U-Net semantic segmentations to YOLOv11 object detectors and zero-shot Foundation

Models, the system accurately detects physical damage to critical survival infrastructure without exposing personnel to ground-level risks. More importantly, by fusing this pixel-level data with vast, unstructured datasets from the ACLED Project, the WFP Logistics Cluster, and local open-source networks, the framework elevates technical damage assessment into actionable, anticipatory intelligence.

We actualize proactive forecasting through application of a structured early-warning typology, identifying structural vulnerabilities, proximate accelerators, and immediate trigger events. We present these data through an intuitively designed, user-centric dashboard prioritizing progressive disclosure, the inverted pyramid layout, and coordinated views. This intelligence empowers policymakers, warfighters, and humanitarian actors to anticipate crises, optimize the allocation of constrained resources, and hold belligerents accountable under international humanitarian law.

Beyond Ukraine and Yemen, our framework offers significant operational utility for monitoring emerging and opaque conflict geographies. GeoAI monitoring could have positive impact in studying emergent grayzone conflict in Iran alongside kinetically dynamic conflicts in Sudan and Gaza. In states with restrictive ground access or suppressed domestic reporting, GeoAI could monitor internal logistical vulnerabilities and regional escalation risk without the need for ground-based sensors or operators.

Ultimately, innovating on the front lines requires more than the deployment of new weaponry; it demands the deployment of new mechanisms for understanding and mitigating the human cost of conflict. By leveraging artificial intelligence, geospatial data fusion, and rigorous ethical guardrails to protect vulnerable populations, this framework not only advances national and global security strategies but also reinforces the fundamental imperative to protect human life and societal resilience in the face of warfare's most destabilizing
realities.

References

[1] Valerie Sticher, Jan D. Wegner, and Birke Pfeifle. Toward the remote monitoring of armed conflicts. *PNAS Nexus*, 2(6):pgad181, 2023. doi: 10.1093/pnasnexus/pgad181. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10291284/.

[2] T. Fernholz. Understanding AI's impact on space data with planet's head of product. *Payload*, 2024. URL https://payloadspace.com/understanding-ais-impact-on-s pace-data-with-planets-head-of-product/.

[3] ICRC. *Protocol Additional to the Geneva Conventions of 12 August 1949, and relating to the Protection of Victims of International Armed Conflicts (Protocol I)*, 1977. URL https://ihl-databases.icrc.org/en/ihl-treaties/api-1977/article-54.

[4] ICRC. *The law of armed conflict at the operational and tactical level*, 1999. URL https://ihl-databases.icrc.org/ru/customary-ihl/v2/rule54. Canada. [Ref: Rule 54 - Attacks against objects indispensable to the survival of the civilian population].

[5] U.S. Department of Defense. *Department of Defense Law of War Manual*, 2023. URL https://media.defense.gov/2023/Jul/31/2003271432/-1/-1/0/DOD-LAW-OF-WAR - MANUAL-JUNE-2015-UPDATED-JULY%202023.PDF.

[6] ICRC International Humanitarian Law Databases. China: Practice relating to rule 53. starvation as a method of warfare, 2025. URL https://ihl-databases.icrc.org/en /customary-ihl/v2/rule53?country=cn.

[7] International Law Commission. Draft principles on the protection of the environment in relation to armed conflict. Technical report, United Nations Office of Legal Affairs, 2022. URL https://legal.un.org/ilc/sessions/73/pdfs/english/poe_eli.pdf.

[8] N. Mohanta. Assessing infra damage in war-torn ukraine, 2022. URL https://geospa tialworld.net/prime/technology-and-innovation/assessing-infra-damage-w ar-torn-ukraine/.

[9] M. Tulbure et al. Leveraging AI multimodal geospatial foundation models for improved near-real-time flood mapping at a global scale. 2024. URL https://www.researchga

te.net/publication/398269597_Leveraging_AI_multimodal_geospatial_foundation_models_for_improved_near-real-time_flood_mapping_at_a_global_scale.

[10] L. Zou et al. Geospatial big data and AI for smart humanitarian mapping, 2023. URL https://www.researchgate.net/publication/398750399_Geospatial_Big_Data_and_AI_for_Smart_Humanitarian_Mapping.

[11] Q. Zhao et al. Satellite and AI monitoring humanitarian crises in the gaza strip during the early stage of israeli–palestinian conflict. *International Journal of Digital Earth*, 2024. URL https://www.tandfonline.com/doi/full/10.1080/17538947.2024.243 0678.

[12] B. Pfeifle. *Detecting armed conflict damages in satellite imagery using deep learning*. Master's thesis, University of Konstanz / ETH Zurich, 2022. URL https://ethz.ch/content/dam/ethz/special-interest/baug/igp/photogrammetry-remote-sensing-dam/documents/pdf/Student_Theses/2022/MA_BirkePfeifle.pdf.

[13] C. Nkolokosa et al. Artificial intelligence-assisted segmentation of flood water from a drone imagery: A use case. *Preprints*, 2025. URL https://www.preprints.org/manuscript/202501.0705.

[14] T. Sweijs and J. Teer. Practices, principles, and promises of conflict early warning systems. Technical report, The Hague Centre for Strategic Studies (HCSS), 2022. URL https://hcss.nl/wp-content/uploads/2022/02/Conflict-Early-Warning-Systems-HCSS-2022.pdf.

[15] ACLED. Armed conflict location and event data project, 2024. URL https://acleddata.com/conflict-data.

[16] A. Voukenas. Mapping the unknown: Using ACLED data and GIS to understand conflict zones, 2023. URL https://medium.com/@avoukenas/mapping-the-unknown -using-acled-data-and-gis-to-understand-conflict-zones-f6f927e04cf1.

[17] ACLED. ACLED conflict alert system (CAST) methodology, 2023. URL https://acleddata.com/sites/default/files/wp-content-archive/uploads/dlm_uploads/2023/07/ACLED-CAST-Methodology-July-2023.pdf.

[18] M. Bertetti et al. Improving the accuracy of food security predictions by integrating conflict data. *arXiv preprint arXiv:2410.22342*, 2024. URL https://arxiv.org/html /2410.22342v1.

[19] WFP Logistics Cluster. Yemen operation: Meeting minutes and access constraints, 2025. URL https://s3.eu-west-1.amazonaws.com/logcluster-web-prod-files/p ublic/2025-05/Logistics%20Cluster_Yemen_Meeting%20Minutes_30042025_Engli sh.pdf.

[20] R. Vos et al. Food crisis risk monitoring: Early warning for early action. Technical report, CGSpace, 2023. URL https://cgspace.cgiar.org/bitstreams/8f3604aa-1 41c-4203-9a77-945a00fc7437/download.

[21] WFP. WFP yemen food security update, december 2025. Technical report, World Food Programme / ReliefWeb, 2025. URL https://reliefweb.int/report/yemen/wfp-y emen-food-security-update-december-2025.

[22] B. P. J. Andree et al. A data-driven approach for early detection of food insecurity in yemen's humanitarian crisis. Technical report, World Bank Policy Research Working Papers, 2024. URL https://openknowledge.worldbank.org/entities/publicati on/b1239319-75d5-4331-ad52-00cf92382111.

[23] B. Bharadwaj et al. Evaluating early-warning systems to improve food security: How to bolster food security through global early-warning systems, 2025. URL https: //www.chathamhouse.org/2025/09/how-bolster-food-security-through-globa l-early-warning-systems/03-evaluating-early-warning.

[24] UN HRMMU. Four years of full-scale invasion of ukraine: Key facts and findings. Technical report, United Nations Human Rights Monitoring Mission in Ukraine via ReliefWeb, 2026. URL https://reliefweb.int/report/ukraine/four-years-ful l-scale-invasion-ukraine-key-facts-and-findings-february-2026.

[25] D. Andrienko et al. Report on damages to infrastructure from the full-scale invasion: November 2024 update. Technical report, KSE Institute, 2025. URL https://kse.ua /wp-content/uploads/2025/02/KSE_Damages_Report-November-2024---ENG.pdf.

[26] UNICEF. The water we leave behind: Securing ukraine's climate resilient future. Technical report, United Nations Children's Fund, 2025. URL https://www.unicef.org/ukraine/en/blog/water-we-leave-behind-securing-ukraines-climate-resilie nt-future.

[27] UNDP. Updated damage assessment finds 524 billion needed for recovery in ukraine, 2025. URL https://www.undp.org/ukraine/press-releases/updated-damage-ass essment-finds-524-billion-needed-recovery-ukraine-over-next-decade.

[28] IPC. Yemen: Acute food insecurity projection update (GoY controlled areas), september 2025 - february 2026. Technical report, Integrated Food Security Phase Classification, 2025. URL https://www.ipcinfo.org/fileadmin/user_upload/ipcinfo/docs /IPC_Yemen_GoY_Acute_Food_Insecurity_May2025_Feb2026_Report.pdf.

[29] Royal Academy of Engineering. Navigating conflict: A comprehensive analysis of water and food security in yemen. Technical report, Food and Agriculture Organization / WANA Institute, 2024. URL https://wanainstitute.org/sites/default/files/publications/%20Final%20Report%20Navigating%20conflict%20water%20and%20f ood%20security%20in%20Yemen_compressed.pdf.

[30] B. Masinde et al. Threat modelling for geodata in the humanitarian context. *Preprints*, 2023. URL https://www.preprints.org/manuscript/202308.0312.

[31] B. Zhao et al. Deep fake geography? when geospatial data encounter artificial intelligence. *Cartography and Geographic Information Science*, 48(4):338–352, 2021. doi: 10.1080/15230406.2021.191007.

[32] S. Praharaj et al. Deploying geospatial visualization dashboards to combat the socioeconomic impacts of COVID-19. *PMC*, 2022. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC9742735/.

[33] UN-GGIM. Geospatial for humanity: Data for action approach, 2021. URL https://ggim.un.org/meetings/GGIM-committee/11th-Session/documents/GeospatialF orHumanity-Covid19.pdf.

[34] A. Rahman. Geospatial dashboards for monitoring smart city performance. *ISPRS International Journal of Geo-Information*, 11(20):5648, 2019. URL https://www.mdpi.com/2071-1050/11/20/5648.

When the Bans Don’t Build Markets: Rethinking the DoD’s Critical Mineral Strategy

Bethany Russell

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Bio

Bethany Russell is a joint MPP and MBA Candidate at Harvard Kennedy School and Harvard Business School, as well as a Research Assistant with the Kennedy School's Defense, Emerging Technology, and Strategy Program. Prior to attending Harvard, Bethany served for six years as an Army Intelligence officer. She can be reached at brussell@mba2026.hbs.edu.

Executive Summary

The U.S. government and the Department of Defense (DoD) are increasingly treating critical minerals as a core national security priority. Over the past several years, both entities have invested heavily in supply-side interventions—financing .projects, supporting processing capacity, and expanding mineral stockpiles. Far less attention however has been given to the design and sequencing of demand-side policies intended to create a durable market for U.S. and allied-sourced minerals.

One of the DoD's most consequential demand-side measures is its 2027 ban on procuring certain covered minerals and components sourced from China. The intent of the ban is clear: reduce dependencies on the People's Republic of China (PRC) and catalyze a secure supply chain. However, as it is currently structured, the ban risks outpacing industrial realities. It imposes a compliance deadline before sufficient alternative mining, refining, and processing capabilities likely exist. It targets minerals that have received comparatively limited DoD investment. It may raise downstream acquisition costs without corresponding budget adjustments. It disproportionately burdens smaller defense firms, potentially narrowing competition within the defense industrial base.

The central argument of this essay is not that demand-side intervention is misguided. Demand-side policies are crucial for the success of the DoD's supply-side investments. Rather, the current approach is poorly synchronized. A ban alone does not create a market; it creates a compliance obligation. Without aligned capital investment, pricing mechanisms, and scale beyond DoD procurement, the ban risks producing bottlenecks, volatility, and consolidation rather than resilience.

To better align demand-side policy with security objectives, the DoD should pursue a compliance framework with five recommendations:

1. **Build a phased compliance framework into the ban.** Rather than enforcing a binary prohibition in 2027, the DoD should add gradual compliance thresholds through using its waiver process. These thresholds can increase over time as verified non-PRC mining, processing, and refining capabilities come online, but they reduce the short-term risk of production shocks and supply bottlenecks.
2. **Incorporate a "mineral security premium" into defense acquisitions.** Secure sourcing will cost more than PRC-derived inputs, and if the DoD does not control for these cost differentials, manufacturers are incentivized to seek waivers to remain price competitive. DoD contracting officers need a formal mechanism to value secure sourcing rather than default to lowest-price comparisons and to reimburse documented price differences associated with complaint sourcing.
3. **Use targeted price-bridging tools, such as contract-for difference mechanisms.** Much of China's mineral advantage comes from scale, state-backed capital, and vertically integrated supply chains, and these overlapping advantages mean it will be difficult for non-PRC suppliers to compete in the short-term. The DoD should therefore deploy mechanisms that temporarily bridge the price gap between PRC and non-PRC minerals, allowing these provisions to sunset as U.S. production reaches scale and cost competitiveness.
4. **Protect and incentivize defense startups.** Mineral compliance regimes impose disproportionate burdens on emerging defense companies, which often lack dedicated compliance teams, long-term supply contracts, or the ability to absorb price volatility. The DoD should establish tiered compliance timelines for small businesses, provide compliance assistance grants or centralized traceability tools, and prioritize startups in transition waivers.
5. **Expand demand coordination beyond the Department of Defense.** The DoD alone does not command sufficient market share to reshape the global mineral demand. To achieve scale, sourcing standards and incentives should be implemented across federal procurement. Coordinated federal demand, coupled with allied alignment and private sector incentives, would create broader and more predictable offtake markets.

Introduction

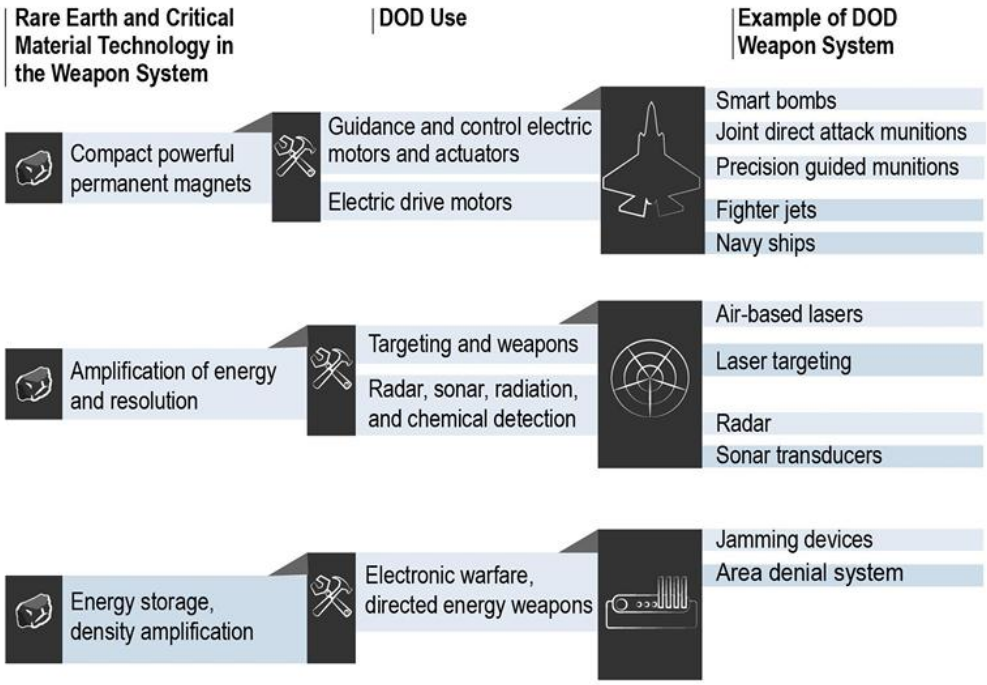
There are more analyses looking at DoD efforts to build supply-side policies promoting critical mineral resiliency, from direct investment to equity. Less attention has been given to DoD's demand-side policies and how it looks to create a market for non-PRC sourced critical minerals. This essay focuses on DoD's efforts to create a mineral consumer base, in particular its ban on defense producers sourcing certain minerals from the PRC after 2027. It explores the motivation driving such a ban, builds a use case from AeroVironment supply chains to model potential impacts for a defense company, and proposes five recommendations for the Department of Defense

Mineral Criticality

The U.S. government broadly defines critical minerals as "nonfuel minerals or materials that are essential to the country's economic or national security and whose supply chain is vulnerable to disruption.[i] This designation reflects two conditions: strategic importance and supply risk. The U.S. government monitors these minerals and their availability closely, and the U.S. Geological Survey must update its list of critical minerals at least every three years, assessing both economic dependence and exposure to geopolitical concentration.[ii]

For the defense sector, criticality is not abstract. These materials possess physical properties that cannot be easily substituted without degrading performance. Rare earth permanent magnets such as neodymium-iron-boron (NdFeB) magnets provide high magnetic strength and can operate under extreme temperatures, making them essential for precision-guided munitions, radar systems, and unmanned platforms. Tantalum's resistance to corrosion and high melting point make it invaluable for aerospace components and missile systems. Tungsten's density and hardness are critical for armor-piercing munitions and high temperature applications.[iii] Figure 1 below from the Government Accountability Office report illustrates some of these potential use cases.

Figure 1. GAO: Defense Uses of Critical Minerals



Source: U.S. Government Accountability Office, "Critical Minerals: Action Needed to Implement Requirements that Reduce Supply Chain Risks," GAO-24-107176 (Washington, D.C.: September 10, 2024), https://www.gao.gov/assets/880/871168.pdf

The vulnerability dimension is equally important. According to the 2025 USGS Mineral Commodity Summaries, China supplies more than 50% of U.S. import demand for over twenty nonfuel mineral commodities, and in several cases it dominates global processing capacity.[iv] Concentration at the mining, refining, or processing stages creates chokepoints. For defense planners, these chokepoints and their disruptions—whether commercial, regulatory, or geopolitical—could constrain the production of key weapons systems. The concentration of these chokepoints in a single geopolitical competitor makes resolving this reliance strategically urgent.

PRC Dominance

China's dominance over critical mineral supply chains is no accident. Instead, it is the product of a multi-decade industrial strategy deliberately designed to target every major node in the value chain. Over the course of two decades, Beijing invested heavily in upstream mining both domestically and internationally. Ultimately the PRC invested over $57 billion in

international mining projects from 2000-2001. These efforts also expanded into refining, processing, and mineral transportation networks, allowing Chinese firms to operate across the entire value chain. As a result, in many mineral categories, raw materials mined outside China are still processed within China before reaching global manufacturers.[v]

This strategy also integrates supply with demand. China's robust domestic manufacturing base produces guaranteed internal offtake. This demand stabilizes domestic producers and insulates them from global volatility. Chinese mineral companies can scale production confidently knowing that downstream buyers exist.[vi] This alignment coordinates capital, production, and consumption into a connected industrial ecosystem, and it grants the PRC structural market power over the global mineral economy.

This market power manifests in two critical ways. First, China's scale allows it to influence global pricing. When Chinese producers expand output, global prices fall; when it restricts exports, prices spike.[vii] Second, the PRC can use strategic oversupply to undermine emerging competitors.[viii] During previous rare earth price cycles, increases in Chinese output contributed to prolonged price depressions that disincentivized non-PRC investment. Private investors and producers in the United States struggle to compete against PRC producers who are reinforced by state financing and embedded in an integrated industrial base.[ix]

For many years, Beijing exercised this dominance cautiously. Aside for 2010 rare earth export restrictions targeting Japan, China largely avoided overt weaponization of its mineral position, relying on more subtle forms of economic statecraft.[x] That strategy changed in 2023, when China imposed export controls on gallium and germanium in response to U.S. semiconductor restrictions. In December 2024, China's Ministry of Commerce added to these bans by prohibiting the export of gallium, germanium, antimony, and graphite,[xi] and the measures continued in 2025 when China banned the export of certain rare earth elements.[xii] These increasing actions suggest the mineral leverage is no longer a theoretical tool but one that China is willing to deploy for geopolitical competition.

The national security implications of Chinese market dominance are self-evident. In a crisis scenario, particular one involving Taiwan, Chinese export controls or licensing slowdowns could constrain U.S. or allied defense production.[xiii] Even in normal peacetime operations, dependence on a concentrated supplier creates vulnerability to price and political leverage.

Defense manufacturers reliant on Chinese inputs are exposed both to price shocks and strategic uncertainty.

These risks are why demand-side policy matters. China's dominance rests on interconnected supply, pricing, and demand scale. Attempting to counter that market dominance through a ban without aligning price mechanisms, investment timing, and broader market demand risks misdiagnosing the source of China's advantage. Competing U.S. efforts therefore need to build coordinated supply and demand policies.

Current DoD Initiatives

Supply Side

In 2022, the U.S. government formally identified fifty minerals as critical to the country's economic and national security, reflecting growing concerns over concentrated global supply chains.[xiv] In the years since, federal supply side measures have accelerated. During the first Trump Administration, a series of executive orders emphasized strengthening domestic mining, refining, production, and recycling capabilities.[xv] The Biden Administration expanded on these efforts, directing significant funding through the Department of Energy's loan programs and the Defense Production Act to support new extraction projects, midstream processing facilities, and battery supply chains across the U.S. and allied countries. It also advanced multilateral coordination through the Minerals Security Partnership and further increased investment in mineral extraction and processing R&D.[xvi] More recently, federal policy has continued to prioritize upstream resilience. Investments in companies like MP Materials, along with initiatives like Project Vault aimed at strategic stockpiling, reflect the ongoing commitment to strengthening non-PRC mineral capacity.[xvii] While specific mechanisms continue to vary, the theme is clear: increase domestic and allied production, diversify supply, and reduce exposure to chokepoints.

Collectively these efforts seek to strengthen the "supply side" of the equation, creating more opportunities for non-PRC mines and processors to enter the market and expand sourcing options for defense and commercial manufacturers. Yet supply expansions alone does not guarantee that demand will materialize in a timely, scalable, and economical manner. This intersection—expanding supply and shaping demand—is where policy becomes more complicated.

Demand Side

While supply-side investments have received the most attention, the Department of Defense has also begun experimenting with demand-side interventions. The DoD has supported select projects through guaranteed offtake agreements and limited price reassurance, reducing the risk level of domestic producers. The DoD's investment in MP Materials, for example, included pricing and purchase commitments to help stabilize rare earth magnet production. More recently, the U.S. government is exploring a broader critical minerals price floor, which would establish a minimum guaranteed price to shield domestic producers from volatility and predatory underpricing.[xviii]

These mechanisms are still evolving, and in scale they remain modest to the scope of the challenge. The most consequential demand-side instrument, however, is regulatory rather than financial: supply chain restrictions. On May 30, 2024, the DoD issued a final rule restricting procurement of certain critical minerals and magnet components from "covered countries", including China, Russia, and North Korea. Effective January 1, 2027, the rule prohibits "samarium-cobalt magnets, tantalum metals and alloys, tungsten metal powder, and tungsten heavy alloy or any finished or semi-finished component containing tungsten heavy alloy." It goes beyond banning minerals mined in China; it includes any mineral "mined, refined, separated, melted, or produced" in one of the covered countries.[xix] This scope is expansive, targeting not only extraction but also processing and transformation stages.

This ban is not absolute. The DoD retains the authority to issue "nonavailability determinations" when compliant suppliers do not exist, allowing temporary exceptions on either an individual or class basis. [xx] Yet because the rule has not taken effect, it remains unclear how accessible these waivers will be in practice or how consistently they will be applied.

By implementing a ban, the DoD is looking to send a powerful demand signal. Although public comments criticized the implementation date (January 1, 2027) for being too aggressive, setting such an ambitious milestone serves as a clear demand signal and timetable to DoD's industry partners.[xxi] This also serves as an informal offtake guarantee. Not only is the government promising to buy non-PRC sourced minerals, but now participants across the defense industrial base will also be sourcing secure minerals. The logic is straightforward: if defense manufacturers must purchase compliant materials, investors can finance new mines and

processors with confidence that buyers exist. This derisks investments in mining, refining, processing, and supply chains. A ban also evens the playing field by preventing firms that use lower-cost PRC inputs from undercutting competitors investing in secure sourcing.

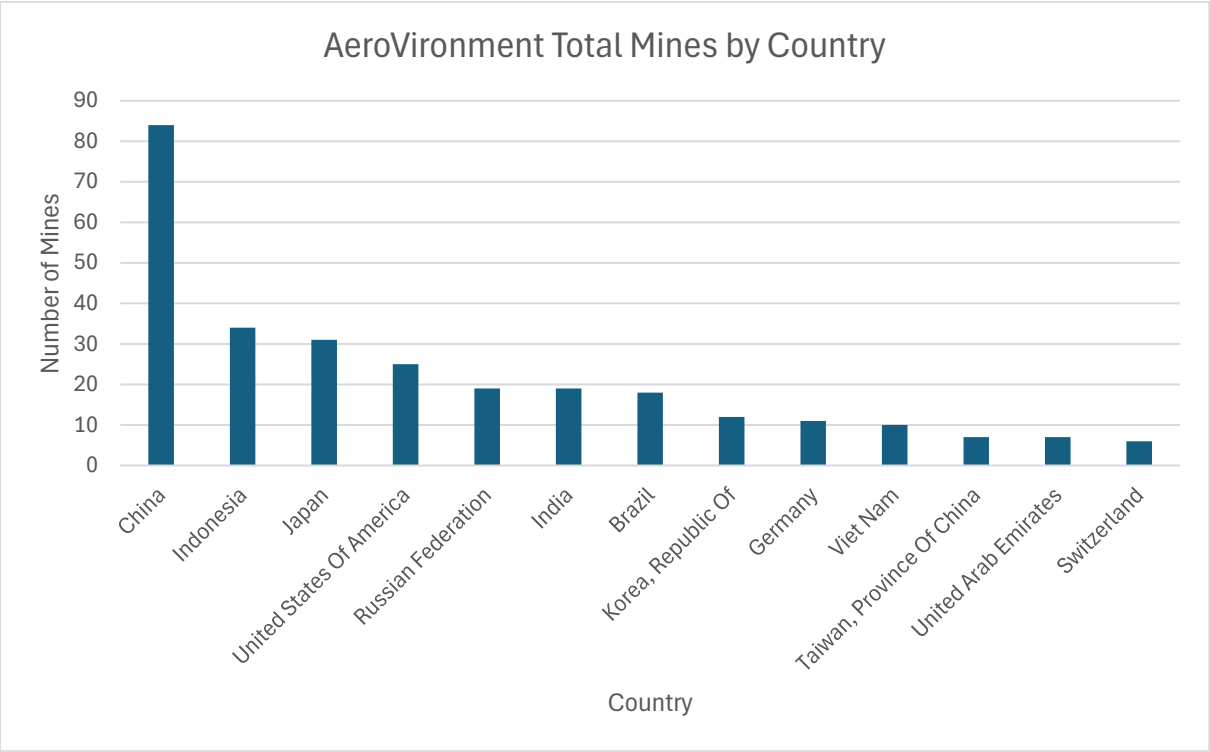
Whether this regulatory demand signal is properly sequenced with supply development, priced into acquisition, or capable of sustaining a competitive industrial base remains deeply uncertain.

Case Study: AeroVironment Mineral Sourcing

To ground this analysis in industrial reality, this paper examines the supply chain of AeroVironment, a U.S. defense manufacturer producing loitering munitions, unmanned systems, and autonomous capabilities for the Department of Defense and allied partners.[xxii] AeroVironment is not uniquely exposed to Chinese mineral supply chains; its sourcing profile likely resembles that of many defense manufacturers. Rather, it is selected for a case study because it publicly discloses select mineral sourcing data through its annual Specialized Disclosure Report filed with the Securities and Exchange Commission (SEC). This provides rare transparency into upstream dependencies.

For its loitering munitions and unmanned systems, AeroVironment tracks four minerals under SEC reporting requirements: tungsten, tantalum, tin, and gold. Across these four minerals, Chinese mines represent the single largest source of supply. In its most recent filing, AeroVironment identified 365 mines across its supplier base, 84 of which are located in China—more than any other country. Indonesia and Japan follow distantly, with 34 and 31 mines respectively.[xxiii] Figure 2 below reflects this supply chain sourcing. Even before examining individual minerals, this distribution demonstrates the broader concentration trends discussed earlier. It's important to note that this analysis relies on mine counts as a proxy for supply exposure. Although production volumes may vary significantly across mines, a lack of mine-level capacity data makes mine count the most reliable available proxy for sourcing concentration.

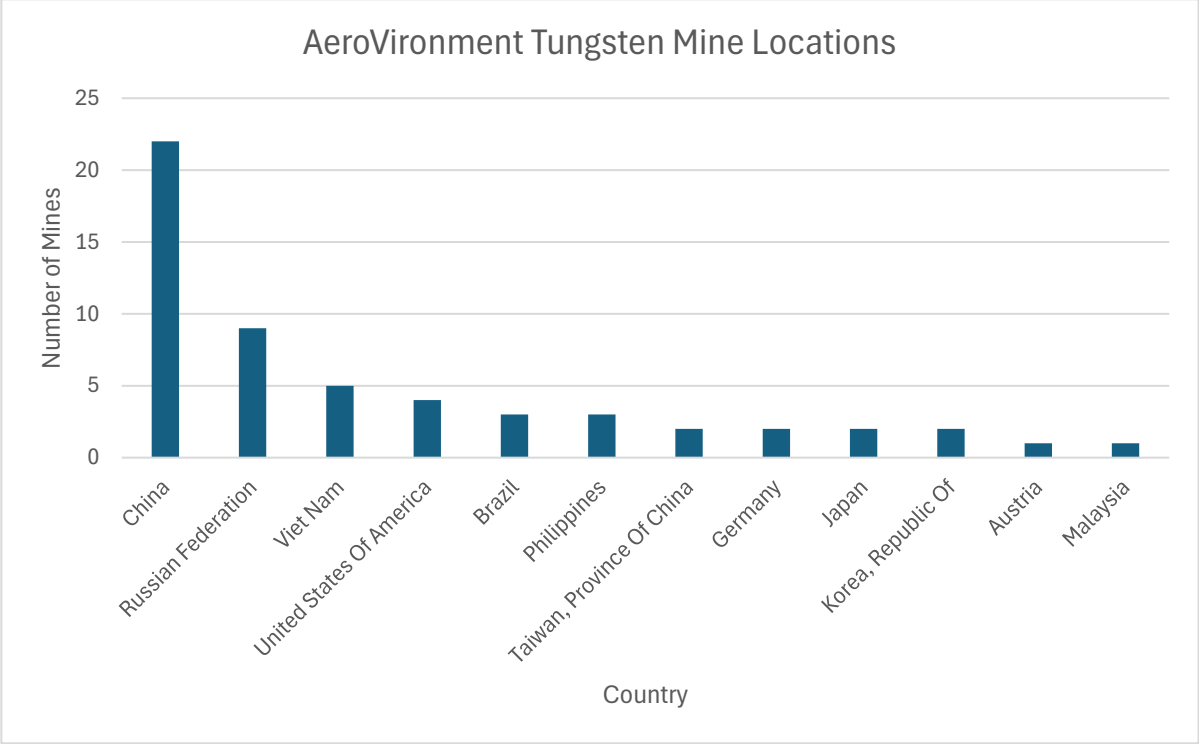
Figure 2. AeroVironment Mines by Country



Source: AeroVironment, Inc., "Specialized Disclosure Report for the Year Ended December 31, 2024," AeroVironment Inc. Note: Chart excludes countries with fewer than 6 mines.

The exposure becomes more pronounced when focusing on the specific minerals targeted by the DoD's 2027 ban. Of the 53 mines supplying tungsten to AeroVironment in 2024, 22 (42%) are located in China. Without Chinese mines, the next largest supplier is Russia—also a prohibited country—with 9 mines (9%). Figure 3 below depicts the extent of this reliance and the availability of alternative suppliers.

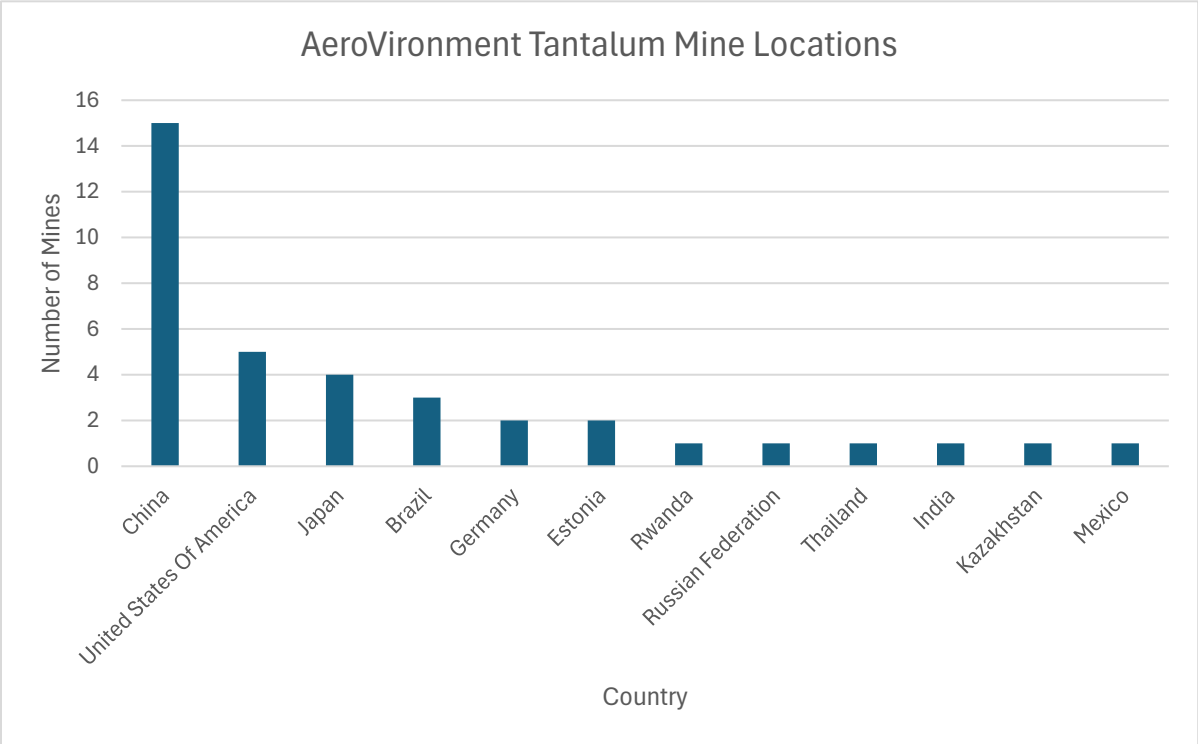
Figure 3. AeroVironment Tungsten Mines by Country



Source: AeroVironment, Inc., "Specialized Disclosure Report for the Year Ended December 31, 2024," AeroVironment Inc.

The supply chain for tantalum, also targeted in the DoD's 2027 mineral ban shows a similar pattern. AeroVironment relies on 36 tantalum mines, 14 (39%) of which are located in China. The next largest supplier, the United States, accounts for only 5 mines (14%). Figure 4 below depicts AeroVironment's tantalum sourcing across its supplier base. These results suggest that nearly two-fifths of AeroVironment's exposure to covered minerals traces directly to China, and that only considers the initial mining location.

Figure 4. AeroVironment Tantalum Mines by Country



Source: AeroVironment, Inc., "Specialized Disclosure Report for the Year Ended December 31, 2024," AeroVironment Inc.

In its SEC disclosures, AeroVironment was transparent about the risks of decoupling its mineral sourcing from China. AeroVironment itself noted that it cannot easily find additional suppliers. The company wrote in its SEC filings that:

> *"Some of our components are sourced from foreign countries which are at risk of sanctions and other trade restrictive actions, such as China, and any escalation in global trade tensions or trade restrictions may hinder our ability to obtain these components from new suppliers. It may take several months to locate alternative suppliers, if required, or to redesign our products to accommodate components from different suppliers. Even if we identify alternate suppliers, we may experience significant delays in manufacturing and shipping our products to customers and incur additional development, manufacturing and other costs to establish such alternative sources, be required to redesign our products and to complete additional quality control procedures."*[xxiv]

The disclosure spells out quite clearly the risks from supply chain disruptions. AeroVironment warns that finding additional suppliers could take several months and may require additional

engineering, quality validation, and manufacturing adjustments. Even after sourcing substitutes, the company anticipates delays in production and shipments, as well as increased development and manufacturing costs. These statements describe risk stemming from geopolitical escalation, but the same operational frictions would likely apply if regulatory requirements abruptly ended PRC-linked supply.

This case study is not an abstract exercise. AeroVironment supplies thousands of Switchblade systems to the war in Ukraine,[xxv] received a $874 million contract in 2025 for allied deliveries,[xxvi] holds a multi-year U.S. Army contract with a $990 million ceiling,[xxvii] and supports Taiwan's asymmetric buildup.[xxviii] Disruptions to its mineral supply chain could affect production timelines for systems currently deployed in active conflict zones and contingency planning environments. The 2027 ban therefore intersects not only with hypothetical supply chains but real-world demand.

Gaps in Current Market Measures

The AeroVironment case helps illustrate the potential for gaps in the current DoD ban on certain minerals. These gaps do not undermine the strategic need for mineral security, but they raise serious questions around if current market interventions are properly sequenced, financed, and calibrated to industrial realities.

Mismatched Timing

Currently the Department of Defense is prioritizing capital behind supply-side measures and for good reason; it faces a supply shortage that is being weaponized by the PRC, and supply policies will take longer to execute. However, its supply and demand interventions are currently mismatched in their timing.

The DoD's ban on certain critical minerals is set to take effect on January 1, 2027, but neither American nor allied capacity is likely to be sufficient to replace PRC supply by that date. New mines can take years to decades to progress from discovery through permitting, development, and production. Industry studies show that mining projects can take well over a decade from discovery to first production, with average lead times increasing as exploration, permitting, financing, and construction phases lengthen.[xxix] Even brownfield sites (existing mines with known deposits) require substantial time to secure permits, build infrastructure, and

scale outputs. The American Tungsten Corp. project in Idaho for example targets a 12-18 month timeline to supply up to 8% of U.S. tungsten demand once operational. [xxx] Even at full projected capacity though, no single project will meaningfully displace U.S. reliance on imports. Closing the gap will require multiple mines reaching production, which compounds capital requirements, permitting timelines, and execution risk.

Beyond extraction, the United States lacks the domestic capacity for intermediate processing and refining of many critical minerals. For tungsten specifically, there has been no commercial mining in the U.S. since 2015, and the country relies on imports and secondary sources.[xxxi] Building integrated processing and transportation networks to move ore through beneficiation and into production will add additional time and capital requirements. These midstream functions mean that even once new mines come online, they will take further time to significantly contribute to defense production.

These factors mean that supply-side initiatives will not align neatly with a 2027 demand restriction. Defense manufacturers facing restrictions on sourcing from the PRC will confront a reality in which there are no compliant alternatives at the capacity and price required. While some intermediate sourcing locations exist, shifting sources to intermediary countries and then again to U.S. or allied suppliers as supply expands will duplicate investments and logistical complexity. In short, the policy timeline moves faster than the industrial timeline, creating compliance obligations that outpace available capacity.

Misaligned Investment

The 2027 ban is narrowly targeted. It applies specifically to "samarium-cobalt magnets, tantalum metals and alloys, tungsten metal powder, and tungsten heavy alloy or any finished or semi-finished component containing tungsten heavy alloy."[xxxii] In other words, the ban focuses on samarium-cobalt magnets, tantalum, and tungsten.

However, recent federal mineral investments emphasize other priorities. The largest single commitment has been the DoD's roughly $400 million funding agreement with MP Materials for rare earth separation and magnet production. Initiatives such as "Project Vault" look to stabilize and expand domestic mineral production more broadly.[xxxiii] While these efforts

address critical vulnerabilities in rare earths and other strategic minerals, they are not tightly concentrated on tungsten and tantalum—the specific minerals listed in the 2027 ban.

Comparatively, DoD funding towards tungsten and related projects has been smaller and often directed towards early-stage work. For example, the DoD awarded Fireweed Metals $15.8 million to advance a tungsten project,[xxxiv] Northcliff Resources received $15 million to accelerate the development of its tungsten efforts,[xxxv] and Golden Metal Resources was awarded $6.2 million for a prefeasibility study of a tungsten project.[xxxvi] These awards are important but modest relative to the scale of U.S. demand. They also primarily target feasibility, testing, and early development rather than output at scale.

This mismatch is meaningful. The 2027 rule focuses regulatory attention on a limited selection of minerals, but capital deployment towards non-PRC capacity for those minerals has been smaller and slower than the deadline otherwise implies. Without greater alignment between the regulatory targets and financing efforts, the burden for compliance will fall downstream on manufacturers.

Downstream Pricing

The ban will also create downstream effects on the cost of goods. Non-PRC suppliers typically face higher labor costs, stricter environmental regulation, and higher energy and compliance expenses than PRC producers. These factors raise the delivered price of processed minerals and components.

Market reporting suggests that the premium for non-PRC materials and components is notable. As an example, estimates from EU manufacturers indicate that a magnet produced in Europe is about 20-30% more expensive than one produced in the PRC.[xxxvii] Taiwanese manufacturers report that drones produced with non-PRC components are on average 25% more expensive that ones that include Chinese components.[xxxviii] These input premiums will not necessarily translate one-for-one into higher end-item prices, given that it depends on the share of the device's value determined by the covered mineral, but they will increase program acquisition costs.

The cost also goes beyond the unit price. Rebuilding vertically integrated mineral transportation networks and supply chains will require new processing capacity, logistics links,

and industrial coordination. These steps add upfront capital expenditures and operating costs that will continue to drive elevated mineral prices until companies achieve economies of scale. Industry cases across the EU and Taiwan demonstrate that these structural costs can be meaningful and persistent unless governments explicitly underwrite them or buyers accept a persistent premium.

A key question emerges around if defense acquisition officers are prepared procedurally and budgetarily to pay more for systems that are identical except for their mineral sourcing. The current ban and waiver mechanisms do not themselves provide a built-in funding mechanism to absorb the premiums paid for non-PRC supply. The recent MP Materials pricing arrangement illustrates that the government can underwrite price stability for secure supply, but it has yet to scale that approach across multiple minerals and programs, particularly against the minerals it is banning.

Discriminatory Impact

The most damaging unintended consequence of a poorly designed mineral ban would be to stifle competition in the defense sector. While the prohibition formally applies to all contractors equally, its real-world impacts will not be evenly distributed. Compliance costs, price volatility, and supply competition disproportionately burden smaller firms and startups—the very actors the DoD has been trying to promote in the defense ecosystem.

Complying with a ban inherently creates compliance costs. It requires supply chain tracing, documentation, auditing, and ongoing monitoring of suppliers. Large defense primes are more able to absorb the costs of regulatory compliance or hire outside advisors to ensure they are in line with DoD policies. Startups are less able to afford outside advisors, meaning they must track their compliance themselves, which can delay production and increase the burden on employees.[xxxix] DoD reporting from other due diligence initiative have noted that in certain conditions, small companies can pay almost twice as much as large ones in conducting self-assessments.[xl]

In addition to administrative burdens, smaller firms are more exposed to mineral price volatilities. In the early stages of transition away from PRC sources, compliant minerals will likely be limited and more expensive. Larger defense primes can more likely secure long-term

contracts, outbid competitors, or vertically integrate to guarantee supply. Startups, who purchase in smaller quantities and operate on tighter margins, may struggle to access compliant minerals at all.

Time horizons also reflect the disparity. Established firms can more easily absorb delays while finding new suppliers or redesigning supply chains for new raw materials. Startups often cannot. Months-long sourcing transitions can jeopardize contract milestones, fundraising rounds, and even survival. Mineral supply uncertainty therefore becomes an existential risk for smaller entrants, potentially discouraging participation in the defense industrial base.

Insufficient Market Creation

A ban does not create a market. What the DoD ultimately needs is a self-sustaining ecosystem in which domestic and allied producers can operate at scale without continuous government intervention. Durable markets require predictable, large-scale demand. This large-scale demand lowers unit costs, attracts private capital, and enables investment in refining, processing, and logistics infrastructure. Without that scale, mineral projects remain dependent on constant policy support rather than commercial viability.

One of China's advantages in mineral competition is its integrated demand. Chinese producers benefit from guaranteed offtake across both the public and private sectors. This constant demand builds confidence in both producers and investors and helps create a virtuous cycle. By comparison, U.S. demand-side policy concentrates primarily around DoD procurement. While Project Vault and other initiatives aim to broaden support, the Pentagon remains the primary driver of mineral policies.[xli]

This scale disparity is significant. The DoD's consumption of many critical minerals represents only a small fraction of global demand. In the case of rare earths for example, DoD demand accounts for less than 0.1% of total global consumption. DoD reports have also noted that commercial markets, not defense ones, drive the development cycles and production capabilities for rare earths and other critical minerals. [xlii] A defense-only prohibition therefore reshapes a narrow slice of total demand without altering global mineral markets.

If the objective is to replicate the economies of scale that underpin China's dominance, demand must extend beyond the Pentagon. A government-wide procurement standard might

expand market signals, but even that may prove insufficient if commercial actors continue to source lower-cost PRC inputs. Without parallel incentives or requirements affecting private sector industrial consumption, a defense-only ban risks creating a high-cost market segment rather than driving structural market changes.

Criticisms

Demand-side policies are not without risk. There are legitimate concerns about cost, urgency, and credibility with a more gradual approach.

First, demand-side mechanisms such as price floors or contract-for-differences are inherently inefficient. If the U.S. government guarantees a price above the global market rate and global prices fall—whether die to cyclical commodity fluctuations or deliberate PRC oversupply—the government must finance the difference. [xliii] In effect, American taxpayers assume the cost of insulating domestic or allied producers from global price volatility. Over time, such investments can become politically vulnerable, particularly if mineral markets stabilize or public attention wavers. While a ban places the compliance obligations on industry instead of fiscal obligations on the government, more complex price-based mechanisms require sustained appropriations and management.

Second, proponents of an aggressive 2027 ban may argue that urgency is of the utmost importance in U.S.-China competition. A hard prohibition forces immediate action. It compels manufacturers to decouple now rather than delay transition decisions until alternate supplies fully emerge on the market. From this perspective, a more gradual mineral transition risks perpetuating dependence: if the DoD gives firms phased targets or generous waivers, they may wait until the last possible minute to adjust supply chains. A strict ban also sends a powerful signal to investors. It demonstrates that the U.S. government firm is willing to reshape markets immediately, creating confidence in the demand for non-PRC minerals.

Third, gradualism can devolve into a permanent transition phase. Political leadership could change, budget priorities shift, and market conditions fluctuate. A phased strategy could become endless extensions, waivers, and delays that combine to mean dependence on PRC minerals never actually lessens. In this view, a firm deadline creates the discipline necessary to achieve security.

These criticisms are valid. However, they do not outweigh the structural risks of a poorly synchronized and implemented ban.

While more demand-side policies like price floors and contract-for-difference mechanisms involve fiscal costs, these costs are transparent and adjustable. By comparison, the costs of a supply shock through a ban, from production delays to startup attrition to higher procurement prices, are diffused throughout the defense industrial base. It is therefore harder to track their impact and monitor the need for change.

Moreover, urgency without capacity creates bottlenecks rather than security. A sudden ban enacted before adequate mining, processing, and refining infrastructure exists will not accelerate production. Instead, manufacturers will compete for non-PRC supply, driving up prices and increasing volatility.

A sudden shock to supply chains also introduces a surge-capacity dilemma. American mineral producers must choose between directing resources towards the Project Vault mineral stockpile or meeting ongoing defense production requirements. A strict ban risks creating competition between U.S. government entities rather than strengthening overall resilience. If stockpile replenishment and current manufacturing efforts draw from the same pool of mineral supplies, the government may end up bidding against itself.

Recommendations

1. Build a phased compliance framework into the ban.

   A sudden ban in 2027 does not reflect the physical realities of mining, refining, and industrial scaling timelines. Rather than an immediate zero-percent threshold for PRC-sourced inputs, the DoD should implement a phased compliance schedule tied to market capacity.

   For example:

   - Year 1: 10-15% minimum sourcing from non-covered countries
   - Year 2: 25% minimum sourcing
   - Year 3: 50% minimum sourcing
   - Year 4: 75% minimum sourcing

A phased structure would accomplish multiple objectives:

1. It provides predictable signals to investors and manufacturers.
2. It reduces shock and disruptions to ongoing defense production efforts.
3. It aligns demand mandates with actual supply development timelines.

Importantly, the thresholds should not be arbitrary. They should be calculated from published assessments of non-PRC mining, processing, and refining capabilities. Compliance targets should only increase as real market availability expands.

2. Incorporate a "mineral security premium" into defense acquisitions.

   If non-PRC minerals cost more, as evidence from European magnet producers and Taiwanese drone manufacturers suggests, these higher input costs must be acknowledged within defense acquisition frameworks. Currently defense manufacturers may face a perverse incentive: seek waivers and source cheaper PRC materials in order to remain price-competitive in defense contract bidding processes.

   The DoD should therefore establish an explicit "secure sourcing premium" within acquisition policy. This could include:

   - Allowing contracting officers to evaluate bids with a standardized mineral security adjustment
   - Publishing modeled price differentials between PRC and non-PRC supplies to better project the implications for input costs
   - Permitting reimbursement of documented mineral security cost differences

   By internalizing the cost of resilience within acquisition policy, the DoD avoids penalizing firms that comply early and rewards those investing in secure supply chains.

3. Use Contract-for-Difference mechanisms to bridge the price gap.

   If the primary barrier to domestic scaling is price competition from subsidized PRC production, then bridging the price gap directly may help diversify sourcing away from the PRC in addition to the outright supply ban.

A contract-for-difference mechanism, established jointly with the Department of Energy, [xliv] could:

- Guarantee domestic processors a fixed price
- Pay the difference between global (PRC-influenced) market price and a secure domestic or allied production price
- Sunset automatically as domestic production reaches cost parity

This model has been used successfully in energy markets to accelerate clean power deployment. Applied to critical mineral markets, it could help stabilize investor expectation and avoid permanent subsidy structures. It could also help bridge the gap between price floors established for American-sourced minerals and PRC-subsidized prices. Contract-for-difference mechanisms would directly address the pricing asymmetry at the heart of PRC market dominance.

4. Protect and incentivize defense startups.

   Compliance burdens and mineral price volatility disproportionately affect smaller firms. Startups, which DoD has sought to encourage through recent acquisition reforms and who are often key drivers of technological innovation, have less working capital, less regulatory capacity, and weaker bargaining power in upstream mineral markets.

   To prevent unintended consolidation within the defense industrial base, the DoD could:

   - Implement tiered compliance timelines for firms below a revenue threshold
   - Provide compliance grants or technical assistance for mineral traceability
   - Establish centralized mineral sourcing databases to reduce reporting burdens
   - Prioritize startups in waiver determinations during transition phases

   If mineral policies inadvertently advantage large incumbers, it will undermine ongoing DoD efforts to diversify and modernize the defense industrial base.

5. Expand demand coordination beyond the Department of Defense.

   The DoD alone does not represent sufficient global demand to reshape mineral markets. Without commercial-scale demand alignment, DoD-specific restrictions risk creating narrow, high-cost enclave markets rather than leading structural changes.

To build real market demand, the DoD could establish:

- Government-wide sourcing standards across federal procurement
- Joint allied sourcing standards within the Minerals Security Partnership
- Broader commercial incentives to drive adoption of non-PRC minerals in the private sector as well

Broader demand increases the certainty of offtake, derisks mineral investments and accelerates economies of scale. Market creation requires scale, and scale requires more than defense procurement alone.

Conclusion

Reducing dependence on Chinese mineral supply chains is a strategic necessity. PRC dominance over mining, process, and refining, paired with its increasing willingness to weaponize this dominance, creates significant leverage in intensifying geopolitical competition. But the current 2027 ban risks moving faster than industrial capacity can adjust. It concentrates regulatory efforts on a limited set of minerals that have received comparatively little investment, may raise acquisition costs, and could disproportionately burden smaller defense firms. A ban can mandate compliance, but it does not generate market resiliency.

---

[i] Diana Roy, "The U.S. Critical Minerals Dilemma: What to Know," Council on Foreign Relations, July 30, 2025. https://www.cfr.org/articles/us-critical-minerals-dilemma-what-know

[ii] Ibid.

[iii] U.S. Government Accountability Office, "Critical Minerals: Action Needed to Implement Requirements that Reduce Supply Chain Risks," GAO-24-107176 (Washington, D.C.: September 10, 2024), https://www.gao.gov/assets/880/871168.pdf

[iv] Roy, "The U.S. Critical Minerals Dilemma: What to Know."

[v] Reed Blakemore and Peter Engelke, *A U.S. Framework for Assessing Risk in Critical Mineral Supply Chains,* Atlantic Council Issue Brief, July 1, 2025. https://www.atlanticcouncil.org/in-depth-research-reports/issue-brief/a-us-framework-for-assessing-risk-in-critical-mineral-supply-chains/

[vi] Ibid.

[vii] Ibid.

[viii] Heidi Crebo-Rediker, "America's Most Dangerous Dependence," *Foreign Affairs,* May 9, 2025. https://www.foreignaffairs.com/united-states/americas-most-dangerous-dependence
[ix] Blakemore and Engelke, *A U.S. Framework for Assessing Risk in Critical Mineral Supply Chains*.
[x] Jonathan Pryke and Nigel Inkster, *Countering China's Coercive Diplomacy,* Australian Strategic Policy Institute, March 2025. https://www.aspi.org.au/report/countering-chinas-coercive-diplomacy/
[xi] Center for Security and Emerging Technology, *"Ministry of Commerce Notice 2024 No. 46: Notice Concerning Strengthening Controls on Exports of Relevant Dual-Use Items to the United States,"* CSET (translation of MOFCOM Notice), December 3, 2024. https://cset.georgetown.edu/publication/china-rare-earth-export-ban/
[xii] Ministry of Commerce, *Announcement No. 18 of 2025: Decision to Implement Export Control on Some Medium and Heavy Rare Earth Related Items*, Ministry of Commerce and General Administration of Customs of the People's Republic of China, April 4, 2025. https://english.mofcom.gov.cn/Policies/AnnouncementsOrders/art/2025/art_0dd87cbee7b045bf93fabe6ab2faceee.html
[xiii] Crebo-Rediker, "America's Most Dangerous Dependence."
[xiv] Roy, "The U.S. Critical Minerals Dilemma: What to Know."
[xv] Crebo-Rediker, "America's Most Dangerous Dependence."
[xvi] Ibid.
[xvii] Gracelin Baskaran and Meredith Schwartz, "Industrial Policy is Back—And Minerals are at the Center," Center for Strategic and International Studies, February 26, 2026. https://www.csis.org/analysis/industrial-policy-back-and-minerals-are-center
[xviii] Ibid.
[xix] "Defense Federal Acquisition Regulation Supplement: Restriction on Certain Metal Products (DFARS Case 2023–D011)," 89 Fed. Reg. 46816, May 30, 2024. https://www.federalregister.gov/documents/2024/05/30/2024-11513/defense-federal-acquisition-regulation-supplement-restriction-on-certain-metal-products-dfars-case
[xx] Ibid.
[xxi] Ibid.
[xxii] "Home," AeroVironment Inc. https://www.avinc.com/
[xxiii] AeroVironment, Inc., "Specialized Disclosure Report for the Year Ended December 31, 2024," AeroVironment Inc, June 2025. https://investor.avinc.com/static-files/7dc8f4fe-91b5-478b-a5fe-3f3d3360b28a
[xxiv] AeroVironment, Inc., *Annual Report on Form ARS for the Year Ended December 31, 2022,* AeroVironment Inc., August 30, 2023, 33. https://www.sec.gov/Archives/edgar/data/1368622/000110465923094840/tm232516d3_ars.pdf
[xxv] AeroVironment, Inc., "Switchblade's Success in Ukraine," AeroVironment Inc. https://www.avinc.com/resources/av-in-the-news/view/switchblades-success-in-ukraine
[xxvi] AeroVironment, Inc., "AeroVironment Awarded $874M Foreign Military Sales IDIQ to Deliver UAS and C-UAS Systems to Allied Partner Forces," AeroVironment, Inc. https://www.avinc.com/resources/press-releases/view/aerovironment-awarded-874m-foreign-military-sales-idiq-to-deliver-uas-and-c-uas-systems-to-allied-partner-forces
[xxvii] AeroVironment, Inc., "AV Secures $288 Million Delivery Order on $990 Million Contract with U.S. Army," AeroVironment, Inc. https://www.avinc.com/resources/press-releases/view/av-secures-288-million-delivery-order-on-990-million-contract-with-u.s-army
[xxviii] AeroVironment, Inc., "AV Signs Strategic Partnership with Taiwan's National Chung-Shan Institute of Science & Technology," AeroVironment, Inc. https://investor.avinc.com/news-releases/news-release-details/av-signs-strategic-partnership-taiwans-national-chung-shan
[xxix] Paul Manalo, "Average Lead Time Almost 18 Years for Mines Started in 2020–23," *S&P Global Market Intelligence*, April 10, 2024, https://www.spglobal.com/market-intelligence/en/news-insights/research/average-lead-time-almost-18-years-for-mines-started-in-2020-23
[xxx] "The American Tungsten Revival Begins in Idaho," *Investor News,* October 14, 2025. https://investornews.com/critical-minerals-rare-earths/the-american-tungsten-revival-begins-in-idaho/

[xxxi] U.S. Geological Survey, *2025 Mineral Commodity Summary: Tungsten*, U.S. Department of the Interior, March 2025, https://pubs.usgs.gov/periodicals/mcs2025/mcs2025-tungsten.pdf
[xxxii] "Defense Federal Acquisition Regulation Supplement: Restriction on Certain Metal Products."
[xxxiii] Baskaran and Schwartz, "Industrial Policy is Back."
[xxxiv] "Department of Defense Makes Investment to Strengthen the Tungsten Supply Chain," Department of Defense, March 7, 2024. https://www.war.gov/News/Releases/Release/Article/4000947/department-of-defense-makes-investment-to-strengthen-the-tungsten-supply-chain/
[xxxv] "Northcliff Announces Funding to Accelerate Development of the Sisson Critical Minerals Project," Northcliff Resources, May 1, 2025. https://www.northcliffresources.com/post/northcliff-announces-funding-to-accelerate-development-of-the-sisson-critical-minerals-project
[xxxvi] "Department of Defense Awards $62 Million to Sustain Critical Production of Tungsten," Department of Defense, October 17, 2025. https://www.war.gov/News/Releases/Release/Article/4252264/department-of-defense-awards-62-million-to-sustain-critical-production-of-tungs/
[xxxvii] Roland Gauß, et al., *Rare Earth Magnets and Motors: A European Call for Action*, The Rare Earth Magnets and Motors Cluster of the European Raw Materials Alliance, Berlin 2021, 11. https://eitrawmaterials.eu/sites/default/files/2024-11/2021_07-13_REE%20Cluster%20Report.pdf
[xxxviii] "Taiwan Flogs America Drones "not made in China," *The Economist,* April 24, 2025, https://www.economist.com/asia/2025/04/24/taiwan-flogs-america-drones-not-made-in-china
[xxxix] Nicole V. Crain and W. Mark Crain, *The Cost of Federal Regulation to the U.S. Economy, Manufacturing and Small Business*, National Association of Manufacturers, October 2023. https://www.nam.org/wp-content/uploads/2023/11/NAM-3731-Crains-Study-R3-V2-FIN.pdf
[xl] *Cybersecurity Maturity Model Certification (CMMC) Program*, 89 Fed. Reg. 83092 (Oct. 15, 2024) https://www.federalregister.gov/documents/2024/10/15/2024-22905/cybersecurity-maturity-model-certification-cmmc-program
[xli] Baskaran and Schwartz, "Industrial Policy is Back."
[xlii] U.S. Government Accountability Office, *Critical Materials: Action Needed to Implement Requirements That Reduce Supply Chain Risks*, GAO-24-107176 (Washington, D.C.: September 10, 2024), 8. https://www.gao.gov/assets/880/871168.pdf
[xliii] Gracelin Baskaran and Meredith Schwartz, *Stabilizing Cobalt Minerals: A Price Floor for U.S. Minerals Security,* Center for Strategic and International Studies, December 10, 2025. https://www.csis.org/analysis/stabilizing-cobalt-markets-price-floor-us-minerals-security#h2-how-price-support-mechanisms-can-support-strategic-minerals-projects
[xliv] Charles Yang, "Demand-Side Financing for Critical Minerals," *The Techno-Industrial Policy Playbook,* 2025. https://www.rebuilding.tech/posts/critical-minerals s

The End of the Gray Zone? How AI-Enabled Cyber Rivals Kinetic Capabilities

March 28, 2026

Daria Bahrami†, Amy Chang†, Michael Kouremetis‡, Erich Devendorf‡, and Tiffany Saade‡

†: Lead authors

‡: Contributing authors

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Bios

Daria Bahrami is Head of Policy at Dreadnode, where she shapes strategic initiatives at the intersection of AI and offensive cybersecurity. Previously, as technical advisor to DARPA's AI Cyber Challenge, she led engagement strategy for global competitions designed to automate vulnerability detection. Daria served as a Lead Cyber Threat Intelligence Analyst at Deloitte Global, briefing executives on emerging threats, and managed critical infrastructure resilience programs. She maintains a Fellowship with the Military Cyber Professionals Association and holds a degree from Georgetown University's School of Foreign Service, specializing in military and space security policy.

Amy Chang leads the AI Threat Research and Security team at Cisco, developing capabilities to secure enterprises from AI-driven risks. She is also Adjunct Faculty in Cybersecurity at the Middlebury Institute of International Studies. With nearly two decades of experience, Amy previously served as Executive Director for Global Cybersecurity Operations at JPMorgan Chase and as a Staff Director for the House Foreign Affairs Committee. A former U.S. Navy officer, her work focuses on scalable frameworks for AI security. Amy is a graduate of Harvard University and Brown University.

Michael Kouremetis is a Principal Engineer at Dreadnode, where he builds offensive cyber agents and evaluations. Previously at MITRE, he led the open-source Caldera project and served as Principal Investigator for advanced R&D programs, including MITRE OCCULT. He has served as a Subject Matter Expert (SME) for autonomous offensive cyber research across DARPA, IARPA, and the DoD. Michael’s work on AI-driven offensive capabilities and adversary emulation has been featured at BlackHat, DefCon, and in Dark Reading, bridging the gap between theoretical models and deployed autonomous systems.

Dr. Erich Devendorf leads the AI Security portfolio at the RAND Center on AI, Security, and Technology. He develops technical and policy options to help ensure the benefits of AI outweigh its risks and harms, drawing on extensive experience in red teaming, offensive cyber operations,

research portfolio planning, and systems engineering. Prior to joining RAND, Erich spent 15 years at the Air Force Research Laboratory, where he created and executed research strategies to maintain technical overmatch against peer rivals.

Tiffany Saadé is an AI security and cyber policy expert working at the intersection of advanced AI systems, adversarial threat intelligence, and national-level governance. She is Product Manager and AI Security Researcher for AI Defense at Cisco, and a Research Associate at the Oxford Cyber and Tech Policy Programme, a Cybersecurity and AI Governance Fellow at the Institute for Security and Technology and a Fellow at Stanford's AI and Future of Decisionmaking in Warfare, where her work shapes global thinking on securing AI systems and integrating AI into cyber defense for critical infrastructure. Tiffany advises the Government of Lebanon nationwide on AI policy and cybersecurity, leads contributions to the country's first national AI framework, and holds a Master's in Cyber Policy & Security from Stanford with a specialization in AI governance and national security. She is one of the youngest people in the Middle East to help design and operationalize a national AI and cybersecurity governance system—at the moment a country is defining its digital future.

Abstract

Historically, cyber operations have failed to deliver the strategic lethality predicted by early theorists. While disruptive, case studies ranging from the Russian power grid attacks to Stuxnet demonstrate that digital effects are often temporary, reversible, and notoriously difficult to synchronize with cross-domain operations. This paper argues that Artificial Intelligence (AI) is fundamentally altering this calculus, collapsing the temporal asymmetry between "gray zone" competition and high-intensity conflict. By automating the speed, scale, and integration of offensive operations, AI provides the "kinetic-equivalent" impact necessary to compel state behavior and alter deterrence dynamics. This analysis employs a conceptual and case-based methodology to explore the trajectory of AI-enabled warfare. First, it identifies the "digital ceiling" of pre-AI cyber operations, analyzing why manual coordination across DIME (Diplomatic, Information, Military, Economic) instruments limited strategic utility. Second, it investigates how agentic AI and Large Language Models (LLMs) bridge the cyber-kinetic gap. Through an examination of autonomous vulnerability exploitation, the research highlights how AI-enabled campaigns can generate continuous, compounding destruction and economic attrition that rivals kinetic bombardment. By collapsing the time between attack and recovery, AI-driven campaigns can outpace an adversary's financial and technical restoration capacity, effectively rendering "reversible" cyber effects strategically permanent. Finally, the paper addresses the critical policy void created by these advancements. Current risk frameworks are static and ill-equipped to measure dynamic AI behaviors, while the democratization of open-weight models renders traditional supply-chain interdiction insufficient. The research concludes that the blurring of the civilian-combatant divide requires a radical shift in U.S. policy: moving from risk aversion to mandatory "resilience by design." This includes establishing minimum security baselines for the private sector and accepting that the democratization of AI capabilities requires a more aggressive posture in offensive testing. Ultimately, this paper posits that to maintain global negotiating power, national security strategy must treat AI-driven cyber conflict not as a support function, but as a primary determinant of geopolitical influence.

1.0 Introduction

For over three decades, cyber operations have occupied a paradoxical position in strategic thought: universally feared, doctrinally prioritized, yet operationally underwhelming on their

own. Despite sustained investment from major powers, no cyber capability has independently achieved the coercive weight, persistent readiness, or decisive battlefield effect that defines a strategic weapon. The 2023 Department of Defense Cyber Strategy conceded the point directly: cyber capabilities held in reserve or employed in isolation render little deterrent effect on their own.[1] Cyber has remained, in the language of conflict theory, a gray zone instrument. This refers to operations that exist below the threshold of armed conflict and are characterized by actions that facilitate espionage, tactical disruption, or economic gain, but are categorically distinct from the kinetic capabilities that compel state-level behavioral change.

This paper argues that artificial intelligence (AI) is eroding that distinction. We contend that AI-enabled automation, including agentic AI workflows, can address the structural deficiencies that have historically prevented cyber weapons from crossing the strategic threshold. These deficiencies in speed, intensity, and control—and related dependencies in scale, persistence, and sustained campaign capacity—are consequences not of the domain itself but of the human operational limitations applied to it.

Several factors exacerbate the current dilemma: first, a pace and capability mismatch between the rapid and continued development of novel generative and agentic AI capabilities and humans' ability to adapt workflows, processes, and policies to account for any novel threats and risks that are introduced. Second, humans do not yet know the upper bounds of generative and agentic capabilities, which risk artificially limiting policies and standards to dictate AI use in conflict. Finally, state-sponsored actors are already integrating AI across offensive cyber campaign lifecycles, from reconnaissance and exploit development through post-exploitation and sustained access, further exacerbating gaps between operations and policy. A new generation of AI benchmarks now quantifies this trajectory with increasing precision, providing a measurement infrastructure capable of tracking offensive cyber capability as rigorously as the defense community tracks missile ranges or warhead yields.

Existing and novel risk frameworks, National Institute of Standards and Technology (NIST) AI Risk Management Framework (AI RMF), MITRE Adversarial Threat Landscape for Artificial-Intelligence Systems (ATLAS), Cisco Integrated AI Security & Safety Framework, and other derivatives, while useful for encapsulating threats and risks that AI technologies introduce, face limitations in accounting for the strategic weight of AI-enabled cyber operations. The policy

implications are urgent and largely unaddressed. International governance efforts, while normatively significant, lack adequate measurement tools, consensus on norms, and any subsequent enforcement mechanisms. U.S. domestic policy infrastructures that operationalize threat assessments of AI-enabled cyber operations into resilience mandates for organizations remain fragmented and underequipped.

This paper proceeds in three parts: Section 2.0 examines the historical record of cyber operations in armed conflict, demonstrating that effective battlefield cyber effects have been limited to conditions of significant overmatch and have failed to achieve strategic impact when belligerents approach parity. It establishes, drawing on classical strategic theory and the Defense Science Board's (DSB) domain-agnostic findings, a three-characteristic framework focused on speed, intensity, and control for what constitutes a strategic cyber weapon. Section 3.0 presents the empirical case that AI is closing the strategic gap, leveraging operational evidence from attributed Russian, Chinese, and Israeli cyber campaigns and analysis of the benchmark data tracking AI capability across targeting, exploitation, and autonomous operations. We propose a velocity-of-attack to velocity-of-recovery (V_a/V_r) framework for assessing cross-domain equivalence between cyber and kinetic effects. Section 4.0 evaluates the policy landscape and highlights the shortcomings of current risk frameworks, the benchmark-to-policy gap, and the limitations of international and multilateral governance structures. The section wraps with a concrete U.S. federal policy proposal centered on machine-readable compliance and open source software resilience as a viable near-term path from threat measurement to mandated defense.

With advancements in generative and agentic AI, the question is no longer whether cyber can achieve strategic weight, but how fast it is doing so, and whether policy can keep pace. The gray zone may be eroding, but the instruments that would measure, govern, and defend against that transformation have not yet arrived.

2.0 The Strategic Gap: Cyber in Armed Conflict

Cyber operations have been employed in armed conflict for over three decades. This section examines the operational record and identifies the structural constraints that have prevented cyber from achieving strategic weight.

2.1 Current State of Cyber Effects in Warfare

Traditional cyber operations, despite their growing sophistication and scale, have historically failed to deliver coercive strategic lethality predicted by earlier theorists. Battlefield effects should not be conflated with intelligence collection or espionage, where cyber has a demonstrated record of success. During the First Gulf War in 1991, U.S. forces were able to operationalize novel technological advancements in electronic warfare to achieve decisive results in paralyzing Iraq's air defense system.[2] Stuxnet disrupted the Iranian nuclear program in 2009.[3] and the 2015 breach of the Office of Personnel Management leaked hundreds of thousands of personnel files.[4] Effective battlefield cyber operations require predictable outcomes, alignment with mission time phasing, and repeatable deployment. Collections and espionage relax these requirements: Stuxnet did not require specific centrifuges to be destroyed at a particular time, and the OPM breach required a single success to achieve its objective. Unless a weapon has enormous destructive potential, repeated employment is required.

Cyber operations have had battlefield impact, but only under conditions of significant overmatch. In the 2008 Russo-Georgia War, the Russian ground offensive was accompanied by cyber effects that likely enhanced the attack and limited the international community's ability to mount a response.[5] In 2016, Operation Glowing Symphony targeted ISIS digital infrastructure as part of the Joint Task Force ARES campaign. However, its effects were not tightly coupled to the daily kinetic operations and were afforded the same latitude in timing and repeatability as espionage actions.[6] Most recently, cyber effects were deployed as part of the U.S. Operation Southern Spear in Venezuela, where Caracas briefly lost power during the ingress of U.S. forces.[7] This operation was executed with timing and predictability, though as a single operation its repeatability remains uncertain.

In all cases, the attackers enjoyed significant capability mismatch with respect to their targets, and cyber was a supporting element to the campaign, rather than a decisive one. In the Russian-Georgia war, the main effort was a rapidly culminating ground invasion. Meanwhile, Southern Spear relied on special operations teams. Cyber effects may have reduced casualties and changed the value proposition for the attacker, but kinetic action was required to achieve the desired end state. Cyber effects alone were insufficient.

2.2 The Human Ceiling: Structural Constraints on Cyber as a Strategic Instrument

When cyber effects have been used during conflicts with closer parity between belligerents, they have had limited impact. During the opening stages of the 2022 Russian invasion of Ukraine, Russia launched a series of cyberattacks designed to cripple Ukraine's communications and energy infrastructure. These attacks were limited in scale during the critical initial phases of the invasion with quick recovery from the damage inflicted.[8] Subsequent Russian cyberattacks have generally failed to achieve impactful effects and have been broadly decoupled from kinetic operations. Failing to generate decisive effects, the preponderance of cyber offense has fallen back to intelligence collections to support waves of cruise missile and drone attacks and information warfare strategies. Even these were observed to "have yet to make a material impact on the battlefield."[9]

The structural limitations described above are not failures of investment or doctrine, but are characteristic of how cyber operations have been conducted for over three decades. The constraints map to human operational capacity: the speed at which analysts develop targets, the rate at which operators generate exploits, the number of simultaneous campaigns a team can sustain, and the tempo at which capabilities regenerate after use.

Breaking through this ceiling requires advances that can generate effects at scale without linear human investment. Delivering those effects requires holding at risk many simultaneous targets and the ability to regenerate combat power at a rate faster than defenders can adapt. All of these advances would have to be sustained over operationally relevant time horizons.

Decisive cyber effects also require something else: the ability to perform battle damage assessments of those effects with sufficient precision to inform strategy and compete for resources. As the cases above illustrate, even successful cyber operations have struggled to demonstrate their value relative to (or as a complement to) kinetic alternatives. It was not because the effects were insignificant, but because there was no accepted methodology for quantifying them. A six-hour grid outage has real economic cost, but without cross-domain equivalence metrics, it registers as "disruption" while a destroyed substation registers as "damage." If cyber has been undervalued partly because its effects resist quantification, then technology that breaks through the ceiling must also provide the metrics that make the

breakthrough legible to policymakers and decisionmakers. The following sections examine both dimensions of this problem.

2.3 Defining the Strategic Threshold

Synthesizing the classical strategic theorists with the Defense Science Board's findings on cyber as a strategic capability,[10] we propose the following definition for what a strategic cyber weapon would entail:

> ***A strategic cyber weapon is a capability that, when initiated, can produce enduring effects against an adversary's vital national interests—at a scale, speed, and reliability sufficient to influence state-level decision-making and alter the strategic balance between nations.***

A strategic cyber weapon is distinguished from tactical cyber tools by three measurable characteristics: speed, intensity, and control. Table 1 defines each.

Characteristics of Strategic Cyber Weapons[11, 12, 13]	
Characteristic	**Determination Criteria**
Speed	**Operational responsiveness**: Sufficiently mature and pre-positioned to generate the desired effect within an operationally relevant timeframe upon decision authority
Intensity	**Strategic-level effects**: Capable of disrupting, degrading, or destroying targets whose compromise threatens national security, economic stability, or societal function (e.g., critical infrastructure, financial systems, military command and control) **Global reach:** Able to access and affect targets at any geographic distance through cyberspace, unconstrained by physical proximity
Control	**Repeatability**: Can be reconstituted and employed repeatedly to sustain operational campaigns, not limited to one-time tactical strikes

	Deterrent credibility: The capability is sufficiently demonstrated or perceived that it alters adversary behavior and strategic calculus at the state level

Table 1: Three characteristics of strategic cyber weapons, measurable determination criteria for classifying cyber capabilities as strategic-level weapons.

2.3.1 What Makes a Conventional Weapon Strategic

In 2018, the Defense Science Board Task Force on Cyber as a Strategic Capability recommended that the Department of Defense "move beyond tactical applications for cyber and realize cyber as a strategic capability."[14] The task force concluded that a strategic capability, regardless of domain or means of employment, must satisfy three generic attributes: it must "create a discernible, and preferably enduring, effect on a target's materiel, efficiency, and/or will"; it must be "sufficiently well developed and mature that it can generate the desired effect within a reasonable time of a stated need"; and it must "be regenerated within a reasonable time" to "support campaigns in addition to one-time [tactical] strikes."[15]

The 2022 Nuclear Posture Review reinforces the relevance of this framework to cyber: it rejected No First Use and Sole Purpose nuclear declaratory policies specifically because of "the range of non-nuclear capabilities being developed and fielded by competitors that could inflict strategic-level damage to the United States and its Allies and partners"[16]—an acknowledgment that capabilities outside the nuclear domain are approaching the threshold of strategic significance.

2.3.2 Why Cyber Weapons Have Not Achieved Strategic Weight

Despite decades of investment and doctrinal ambition, cyber weapons alone have not crossed the strategic threshold defined above. The 2023 Department of Defense Cyber Strategy conceded the point directly: "Cyber capabilities held in reserve or employed in isolation render little deterrent effect on their own."[17] The 2017 DSB Task Force on Cyber Deterrence was equally blunt: "Unlike precision-guided munitions, cyber weapons cannot be bought and deployed on a delivery system (or placed in a storage site) with confidence that they will work when needed."[18]

These discrepancies are both structural and technical: cyber's uncertain deterrence value and variable effects fundamentally distinguish it from the nuclear domain. Nuclear weapons gain deterrent value from being advertised, while cyber weapons and "offensive cyber operations are better used than threatened."[19] Even when employed, cyber weapons alone also have "limited effectiveness as an independent tool of coercion,"[20] and Libicki argues similarly that cyberwar faces limitations: "it cannot disarm or destroy the enemy, and …cannot lead to territorial conquest."[21] On the technical side, cyber weapons require a killchain that has distinct phases between acquiring access to a target and executing a cyber effect against a target, which introduces additional dependencies and considerations at each step; Dykstra et al. further noted that cyber weapons are sensitive to environmental changes and can be copied and reused by adversaries.[22] Maschmeyer formalized these technical constraints as the "subversive trilemma," where speed, intensity, and control are negatively correlated, meaning "a gain in one variable tends to produce losses across the other two."[23]

We argue that six deficiencies have prevented current cyber weapons from satisfying the three strategic characteristics (speed, intensity, control):

Deficiencies Preventing Cyber Strategic Weight[24, 25, 26, 27, 28, 29]		
Deficiency	**Description**	**Strategic Deficiency**
Signaling Paradox	Revealing a cyber capability to deter an adversary simultaneously enables the adversary to patch, reconfigure, or neutralize the threat. "Cyber weapons are better used than threatened; nuclear weapons are better threatened than used."	Control
Environmental Fragility	Cyber effects are extremely sensitive to changes in the target's software, hardware, or user settings. Because cyber weapons require establishing and maintaining covert access long before effects can be delivered, small unobserved changes to the target environment can sever the access path and render the weapon ineffective. [1][15]	Speed, Intensity, Control

Adversary Replicability	Unlike kinetic weapons, which are typically destroyed upon employment, cyber weapons can be observed, copied, and reused by adversaries, creating a “glass house” dynamic that constrains employment.	Control
The Subversive Trilemma	Speed, intensity, and control are negatively correlated in cyber operations. Increasing one degrades the others, meaning cyber weapons are rendered “too slow, too weak, or too volatile” to produce reliable strategic effects.	Speed, Intensity, Control
Inability to Disarm or Destroy	Cyber operations cannot permanently disarm an adversary's military forces, destroy war-making capacity, or compel territorial concessions. Effects are reversible and transient, lacking the irreversibility that defines strategic-level damage.	Intensity
Coercive Inadequacy	Cyber power alone fails as a tool of independent coercion. Of six warfighting strategies tested against cyber (attrition, denial, decapitation, intimidation, punishment, risk), most prove ineffective, and deterrence frameworks developed for nuclear weapons are structurally unsuitable for the domain.	Control

Table 2: Six deficiencies inhibiting cyber strategic weight, mapped against strategic characteristics.

3.0 AI Closes the Gap

This section presents the theoretical framework, operational evidence, and quantitative benchmarks demonstrating that AI is systematically closing long-held assessments about cyber as a strategic weapon.

3.1 How AI Can Close the Strategic Gap

The six deficiencies above are not immutable laws of the domain; rather, they are consequences of human operational limitations applied to a uniquely complex environment. AI enablement addresses these constraints directly by providing scale across all key components of cyber operations simultaneously: targeting, vulnerability discovery, exploit development, access

maintenance, effect delivery, and operational adaptation. The result is the potential for on-demand, reliable, wide-scale effects against significant singular or collective targets, pushing cyber capabilities into the sphere of strategic weapons. The following sections examine the empirical evidence for this transformation; here, we outline the theoretical mechanism by which AI addresses each deficiency.

Signaling Paradox. The signaling paradox assumes a finite set of capabilities that, once disclosed, are exhausted. AI alters this calculus through volume and regeneration. A system that continuously discovers new vulnerabilities, generates novel exploit variants, and develops alternative access paths renders the disclosure of any single capability non-fatal to the overall threat. The strategic weapon is not any individual exploit; it is the system's demonstrated capacity to guarantee effects through one of n methods, where n is continuously replenished. Deterrent credibility follows not from concealing a specific capability, but from demonstrating an inexhaustible capacity to generate new ones.

Environmental Fragility. An always-on AI system tasked with maintaining operational readiness against a target transforms fragility from a point-in-time vulnerability into a continuously managed process. Rather than pre-positioning a static capability and hoping the environment remains stable, the system monitors target configurations, adapts to changes in real time, and maintains multiple concurrent access paths. The result is functional equivalence to the persistent readiness that characterizes traditional strategic weapons: the capability to deliver effects on demand, regardless of when the order comes.

Adversary Replicability. A captured exploit or malware sample reveals one instantiation of the weapon; it does not transfer the system's capacity to generate alternatives, adapt to novel environments, or coordinate across multiple simultaneous operations. Many cyber capabilities are uniquely tailored to exploit specific vulnerabilities in an adversary's systems. By nature of their bespoke design, they cannot be repurposed since the weapon only fits one target. The strategic component of an AI-enabled cyber weapon is the autonomous system that guarantees an effect through one of many possible methods—not the specific method employed in any given instance.

The Subversive Trilemma. Maschmeyer's trilemma exists because the complexity of simultaneously maintaining speed, intensity, and control exceeds human operational capacity.[30] An AI system compresses that trade-off space when it autonomously adapts to target environments in real time, generates novel exploit variants dynamically, and maintains precise targeting across multiple simultaneous operations. Whether AI fully breaks the trilemma or merely shifts the frontier outward is an empirical question examined in later sections, but the direction is clear: machine-speed adaptation reduces the penalty for pursuing speed and intensity together, while autonomous targeting preserves control that human operators would sacrifice under time pressure.

Inability to Disarm or Destroy. The claim that cyber operations produce only reversible, transient effects understates the range of achievable outcomes. Data, intelligence, financial records, intellectual property, and operational plans can all be permanently deleted or corrupted. Sometimes deleting data functionally destroys the system, as with wipers and ransomware. Sustained denial-of-service against critical systems, maintained over operationally relevant time horizons by an autonomous system, produces effects functionally equivalent to physical destruction from the target's perspective: a power grid held offline for weeks or a financial system rendered unreliable for months imposes costs comparable to kinetic damage, even if the underlying infrastructure remains physically intact. AI extends the duration and reliability of these effects by automating the sustained campaign that human operators cannot maintain at scale.

Coercive Inadequacy. Coercive inadequacy is a downstream consequence of the preceding five deficiencies—cyber has failed as a coercive instrument because it has been too unreliable, too fragile, and too impermanent to compel changes in adversary behavior. To the extent that AI automation resolves the first five deficiencies, coercive adequacy follows.

A necessary counterargument is that AI scales defense as well as offense, where the same autonomous vulnerability discovery that enables exploit generation also enables patching at machine speed, as AI Cyber Challenge and Big Sleep have demonstrated on the defensive side. Defenders could similarly deploy AI agents to monitor, detect, and remediate at the same tempo attackers operate, which would ultimately rebalance the broader strategic calculus.[31,32]

For the time being, the structural asymmetry of AI adoption and use favors offense for three reasons. First, the attacker chooses the time, target, and method, while the defender must continuously protect all elements. AI amplifies this asymmetry because an attacker needs one successful path, while the defender must maintain adequate resilience against each potential compromise. Second, the deployment incentives are misaligned: offensive AI can be fine-tuned, run unrestricted, and deployed without audit trails on self-hosted infrastructure, while defensive AI operates under compliance, procurement, and integration constraints that slow adoption, particularly across the fragmented civilian infrastructure that constitutes the actual attack surface. Finally, institutions have yet to establish norms or policies for how AI materially alters the cyber defense landscape and best practices to address the potential ecosystem of threats.

3.2 Operational Evidence

The theoretical mechanisms described in Section 3.1 generate specific empirical predictions: if AI technologies are closing the strategic gap, their integration into offensive cyber operations across reconnaissance, exploitation, and sustained campaigns should be at least partially observable and should satisfy the three strategic characteristics defined above. Derived from our open source intelligence assessments, this section examines specific attributed cases of AI integration into cyber operations. It is organized by capability phase to parallel the strategic framework, progressing from targeting through exploit development to full campaign operations. Section 3.3 then examines a complementary question: whether the emerging landscape of AI benchmarks can quantify the trajectory of these capabilities with sufficient rigor to inform policy.

3.2.1 Cyber Reconnaissance and Targeting

AI-enabled reconnaissance and targeting have demonstrated proficiency in the cyber domain. In February 2024, OpenAI and Microsoft disclosed they had disrupted state-affiliated threat actors from Russia, Iran, and North Korea attempting to leverage ChatGPT to support cyber operations, including open source intelligence gathering and target research.[33] In November 2025, Anthropic reported a Chinese state-sponsored actor GTG-1002 leveraged its chatbot Claude to support "reconnaissance, vulnerability discovery, exploitation, lateral movement, credential harvesting, data analysis, and exfiltration operations."[34] Anthropic's Claude was also reportedly deployed on

classified Pentagon networks via Palantir during the January 2026 operation against Venezuelan President Maduro, though the specific role AI played in targeting or reconnaissance for that operation remains unclear.[35,36]

Academic and practitioner research has confirmed the scale: large language models (LLMs) can automate the full reconnaissance-to-attack pipeline of spear phishing campaigns, with AI-automated attacks matching human expert performance at a 54% click-through rate—a 350% improvement over generic phishing with no human in the loop.[37,38] These capabilities represent progression towards strategic-level cyber reconnaissance when compared to traditional targeting methods. The acceleration is structurally analogous to what Israel demonstrated through kinetic targeting during military operations in Gaza, where Israeli officials stated that after an 11-day siege in Gaza in May 2021, it had fought its "first AI war."[39] Former Israeli Defense Forces (IDF) Chief of Staff stated that in Operation Guardian of the Walls in 2021, the IDF's use of AI compressed target generation from approximately 50 per year to 100 per day.[40] After target selection, Israeli officials stated they would use another AI model, Fire Factory, "to calculate munition loads, prioritize and assign thousands of targets to aircraft and drones, and propose a schedule."[41] This enhanced ability to pair sensors with shooters in the cyber domain demonstrates effects at scale, serving to resolve the problem of control outlined in the subversive trilemma. Articles written by IDF officers referencing recent Israeli military campaigns have explicitly acknowledged that "AI is no longer merely a tool but a strategic capability on par with airpower or nuclear deterrence."[42]

3.2.2 Malware & Exploit Development

While not directly from the battlefield, the introduction of more capable models and agentic systems demonstrated progress toward AI systems capable of reliably producing on-demand exploits and malware across diverse target environments. Google's Big Sleep agent, a collaboration between Project Zero and DeepMind, discovered a previously unknown memory-safety vulnerability in a development version of SQLite that 150 CPU hours of traditional fuzzing failed to find.[43] By July 2025, Big Sleep had identified CVE-2025-6965, a separate SQLite flaw known only to threat actors and imminently planned for exploitation, compressing the entire detection-to-patch lifecycle to 48 hours.[44]

Defense Advanced Research Projects Agency (DARPA)'s AI Cyber Challenge (AIxCC) Final Competition at DEF CON 2025 demonstrated autonomous cyber reasoning systems identifying 86% of synthetic vulnerabilities across 54 million lines of critical infrastructure code, up from 37% just one year earlier, and successfully patching 68% of total synthetic vulnerabilities presented.[45] Anthropic reported a comparable result in February 2026: Claude Opus 4.6, with no task-specific tooling or custom scaffolding, discovered and validated over 500 high-severity vulnerabilities across major open source projects by reasoning about code logic rather than relying on brute-force input generation, even in codebases with millions of CPU hours of prior fuzzing.[46]

On the malware development side, Google's Threat Intelligence Group in November 2025 documented five novel AI-powered malware families. Among them are PROMPTFLUX, which uses the Gemini API to rewrite its own malware code hourly to evade detection, and PROMPTLOCK, an experimental ransomware generator that dynamically crafts malicious scripts by querying LLMs at runtime.[47] Academic research has confirmed the feasibility at scale: LLMalMorph generated 618 functional malware variants from ten Windows malware samples without any model fine-tuning, reducing antivirus detection rates by 10–15% and achieving up to 91% evasion rates against machine-learning-based detectors.[48]

Taken together, these developments indicate that the capability to autonomously discover vulnerabilities, generate working exploits, and produce detection-resistant malware variants is no longer theoretical; it is being demonstrated in controlled environments at near-operational fidelity, and in several cases has already crossed into live deployment. As AI capabilities advance, the speed constraints inherent in the subversive trilemma are eroding, with the timelines for developing and executing implants, payloads, exploits, and operations shrinking rapidly.

3.2.3 Campaigns & Post-Exploitation Operations

No complete example of AI-enabled strategic cyber campaigns exists in the open source literature, but nascent, attributed examples demonstrate how AI actively pervades more phases of cyber operations. For example, the U.S. Department of Defense and the intelligence community have reported in their 2025 Annual Report to Congress on the Military and Security Developments Involving the People's Republic of China that commercial AI developed in China

"probably will benefit the [People's Liberation Army's (PLA)] cyberspace capabilities" in both offensive and defensive operations.[49] This section elaborates on several attributed cyber operations associated with Russia, China, and Israeli actors.

Russia: AI as a Tactical Tool

Russian state-sponsored use of AI for offensive operations is most apparent in the information warfare domain, but there is also evidence they leverage AI for tactical cyber operations.[50] In July 2025, Ukraine's Computer Emergency Response Team identified LAMEHUG (also known as PROMPTSTEAL) malware family associated with APT28 (Russian foreign military intelligence unit GRU) that integrated AI into its workflow to dynamically generate commands to enhance or obfuscate Russian cyber activity.[51,52] In the reconnaissance phase, PROMPTSTEAL queries an open source model (Qwen2.5-Coder-32B via Hugging Face API) to dynamically generate system enumeration commands, replacing static collection scripts with AI-driven target mapping that adapts to each environment.[53] During collection and exfiltration, the LLM identifies which documents are worth stealing rather than following a predetermined list, meaning the malware effectively triages intelligence value in real time.[54,55] For persistence and evasion, a related family (PROMPTFLUX) dynamically obfuscates its own code and generates malicious functions on demand, producing polymorphic payloads resistant to signature-based detection.[56,57] Across the campaign lifecycle, Ukraine's State Service for Special Communications and Information Protection confirmed that broader Russian malware samples showed "clear signs of being generated with AI," extending beyond phishing lure creation into the malware code itself.[58,59]

China: AI as Operational Infrastructure

Anthropic's threat intelligence team identified a "sophisticated Chinese threat actor" that systematically leveraged Claude across 12 of 14 MITRE ATT&CK tactics over a nine-month campaign targeting Vietnamese critical infrastructure.[60] The actor compromised major Vietnamese telecommunications providers, government databases, and agricultural management systems in what Anthropic assessed as an intelligence collection operation.[61] Claude was integrated as technical advisor, code developer, and operational consultant throughout the attack lifecycle: building custom Python scanning tools for reconnaissance of Vietnamese IP ranges,

creating file upload fuzzing tools and WordPress exploitation frameworks, optimizing credential harvesting with tools like Hydra and hashcat, implementing Linux kernel privilege escalation exploits, configuring proxy chains for operational security, and analyzing reconnaissance data to plan lateral movement. The case demonstrates a China-nexus actor embedding AI across nearly every phase of a sustained cyber campaign as persistent operational infrastructure.

Israel: Approaching Strategic Posture

Open source intelligence and reporting indicate that Israel's recent cyber-kinetic operational record is the strongest publicly documented example of the progression of cyber as a strategic capability. In short, Israel has shown the ability to maintain persistent access across multiple Iranian infrastructure sectors—including fuel distribution, steel manufacturing, military banking, cryptocurrency exchanges, and broadcast satellites—and activates these capabilities both as technical conditions allow and political imperatives arise. What makes this posture strategically viable long-term is the integration of AI capabilities. Maintaining dormant access across this many sectors simultaneously (i.e., tracking configuration changes, credential rotations, and patch cycles that could invalidate any foothold) exceeds human monitoring capacity. Unit 8200's hundred-billion-word Arabic LLM enables continuous intelligence processing across all targets at once,[62] while AI targeting systems have markedly improved target generation and fire preparation.[63,64,65] These capabilities were on display in June 2025 during Israel's Operation Rising Lion, a 12-day military campaign against Iran. Four days after airstrikes on Iran began on June 13, observers noted synchronized cyber effects across banking, cryptocurrency, and broadcast infrastructure.[66,67] After the 12-day campaign, reports acknowledged the "indispensable" role of AI in battlefield operations.[68] The report detailed the fusion of AI systems with cyber intelligence feeds to "deliver high-resolution, real-time situational awareness" at "speeds impossible under traditional command structures."[69] AI has transformed cyber from a capability that must be painstakingly rebuilt for each operation into something resembling a traditional strategic weapon: persistently deployed, continuously maintained, and deliverable at the speed of political decision.

Table 3 below highlights major cyber-kinetic operations that correlate to political triggers rather than technical windows, affirming these capabilities are used at will and are not solely

opportunistic. This distinguishes them from traditional offensive cyberattacks and marks them as strategic cyber weapons.

(Select) Israeli Cyber Operations Against Iran, 2021–2025 [70, 71, 72, 73, 74, 75, 76, 77]			
Operation	**Effect**	**Political Trigger**	**Strategic Weight Qualities**
July 2021 Iranian Railway System	Nationwide rail paralysis; Ministry of Roads disrupted	Covert pressure campaign against regime	• Meteor wiper selectively avoided PIS display servers (checked hostnames) to preserve messaging channels while destroying backend systems—controlled, deterministic targeting logic • Demonstrated on-demand reach into national transport infrastructure
October 2021 Iranian Gas Stations (1st Strike)	~80% of 4,300+ stations disabled; payment systems destroyed	Anniversary of Nov 2019 protest crackdown; proxy escalation	• Custom Meteor wiper pre-built for this target; deployed deterministically at chosen time for predictable nationwide effect • Access to semi-air-gapped National Information Network indicates deep pre-positioning • Attackers chose which stations to spare and warned emergency services—calibrated yield control
June 2022 Khuzestan Steel Mill	SCADA compromised; molten steel spill causing fire; two additional steel	Escalation after Iran's 2020 cyberattack on Israeli water infrastructure	• Crossed cyber-physical threshold—digitally caused kinetic destruction—joining only Stuxnet and the 2014 German steel mill

	companies hit simultaneously		• Compromised Siemens PCS7/S7-400 controllers requiring deep industrial process knowledge • Used real-time CCTV to verify floor was clear before triggering—risk-assessment protocol consistent with regulated military operations • Three dispersed targets hit simultaneously
December 2023 Iranian Gas Stations (2nd Strike)	~70% of stations disabled again; nationwide fuel chaos (Picus Security, 2025)	2 months post-Oct 7; retaliation for proxy aggression	• Re-strike against a target Iran had 2 years to harden—cyber equivalent of second-strike capability • Group stated it could disable all stations but chose not to—scalable yield • Held in reserve and deployed at a strategically chosen moment
June 17, 2025 Bank Sepah (IRGC Bank) Data Destruction	Bank data reportedly erased; IRGC payroll disrupted; ATMs dark; branch closures (Lyngaas, 2025; TechCrunch, 2025)	4 days after Israeli airstrikes on nuclear facilities	• Timed to coincide with kinetic airstrikes—integrated cyber-kinetic planning • Mandiant's Hultquist: actor is 'not all bluster'; former NSA cyber director Joyce confirmed 'tangible effects' • WSJ reported deep pre-positioned access • Coordinated with Nobitex strike next day.
June 18, 2025 Nobitex Crypto Exchange	$90M drained and destroyed via burn addresses; source code and docs leaked; platform inoperable	Coordinated with Bank Sepah attack previous day	• Funds deliberately burned, not stolen—purely destructive military objective, not criminal. • TRM Labs found VIP compliance-bypass logic identifying IRGC accounts—requires years of deep access

($90M Burned)			

Table 3: Selected Israeli cyber operations against Iran, including operations, effect, political triggers, and qualities associated with strategic weight.

Table 4 below maps the three cases against the five strategic characteristics defined in Section 2.1.1, illustrating a progression from tactical augmentation through operational integration toward strategic posture.

Assessment of Strategic Weapons Criteria Against Observed Cyber Campaigns			
Strategic Weapon Characteristic	**State-Nexus Actor**		
	Russia (APT28 LAMEHUG Campaign)	**People's Republic of China (Vietnam Campaign)**	**Israel (Iran Operations)**
Speed	*Not observed*	Emerging: multi-month persistent access	Demonstrated: activation on political timeline
Intensity	Emerging: adaptive reconnaissance, polymorphic-like evasion	Emerging: 9-month sustained access to victim environment; use of AI across 12 out of 14 MITRE ATT&CK tactics	Demonstrated: multi-sector (banking, broadcast, fuel, manufacturing) simultaneous effects
Control	*Not observed*	*Not observed*	Emerging: escalatory patterns observed, no

			confirmed behavioral change

Demonstrated = operationally confirmed | **Emerging** = partial or evolving evidence | **Not observed** = no public evidence

Table 4: Assessment of Strategic Weapons Criteria Against Observed Cyber Campaigns.

The evidence above demonstrates that AI is already permeating offensive cyber operations across the full campaign lifecycle: from tactical augmentation (Russia) to operational infrastructure (China) to what may approximate strategic posture (Israel). But merely establishing that these capabilities exist is not the same as measuring their trajectory. A claim of strategic significance—one that should inform deterrence posture, resource allocation, and arms control—requires more than operational anecdotes. It requires quantitative tools that can track how fast capabilities are improving, benchmark them against meaningful thresholds, and translate the results into language that policymakers can leverage. A new generation of AI evaluations is emerging to provide exactly that.

3.3 Measuring the Trajectory: AI Benchmarks as Proxies for Offensive Cyber Capability

The operational evidence in Section 3.2 establishes that AI is already integrated into offensive cyber campaigns. This section addresses a complementary and equally consequential question: Can the trajectory of AI-enabled offensive capability be measured with sufficient rigor to inform strategy? New AI benchmarks designed to evaluate reasoning, tool use, code comprehension, and sustained autonomous behavior serve as increasingly reliable proxies for the offensive cyber skill set. Their value is dual: they demonstrate how fast capabilities are advancing, and they constitute the first quantitative infrastructure that could underpin a measurement regime for AI-enabled cyber threats. The following data in Figure 1 below should be read through both lenses.

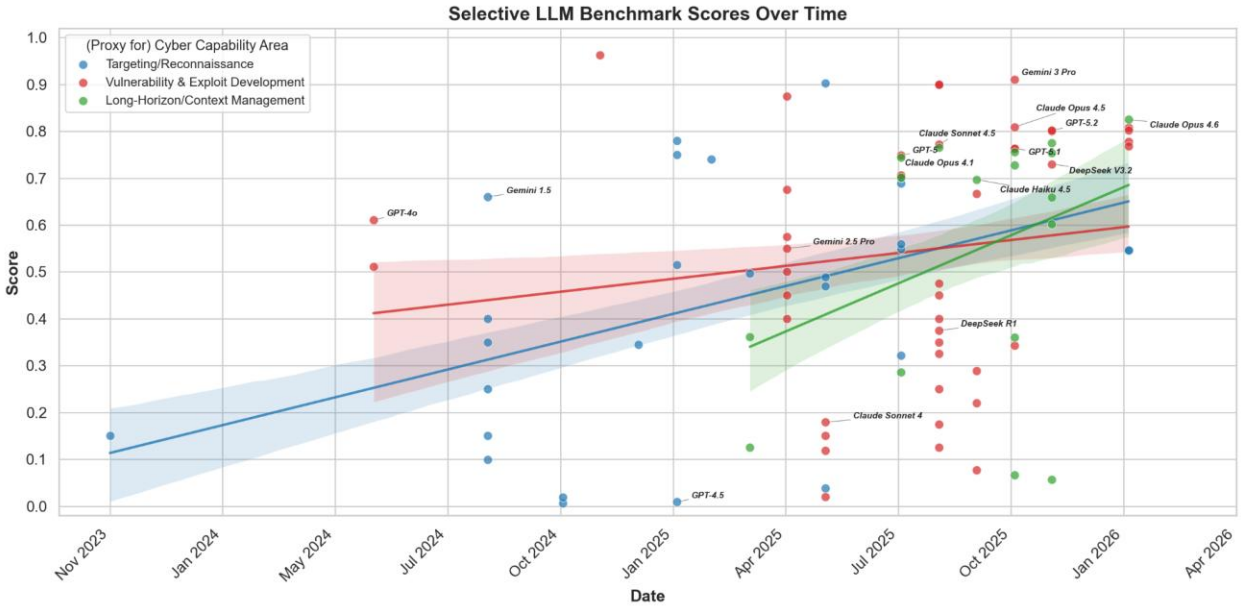


Figure 1. Selective LLM benchmarks (Nov 2023–Feb 2026) across capability areas that can serve as proxies for offensive cyber capability areas. Scores are normalized to 0–1 across heterogeneous metrics. Select model families are labeled to highlight generational improvement. Confidence bands are one standard deviation. Refer to Appendix A for more details about our benchmark assessments.

Advancements across targeting, exploitation, sustained operations, and cost efficiency are compounding simultaneously, marking a convergent leap rather than incremental progress in any single domain. Model Evaluation and Threat Research's time-horizon doubling rate—roughly every seven months—serves as a predictive scaling law for autonomous agent capability, suggesting full cyber campaigns could soon unfold within days.[78] Cost trends reinforce this acceleration: model efficiency, quantization, and distillation are driving exponential reductions in compute and deployment expenses. As training, inference, and large-scale data processing costs plummet, even modest budgets can now sustainably operate large numbers of autonomous agents without oversight. Combined with the democratization of complex cyber capabilities being increasingly available through commercial AI models, these developments transform AI-enabled cyber operations from a theoretical concern into a measurable strategic threat capable of continuous, coordinated action at scale.

3.4 Reversibility No More: How AI Challenges Permanence

A central tenet of cyber conflict theory is that the effects of cyberattacks are reversible: systems can be rebooted, code patched, and data restored from backups.[79] This assumption has anchored the conventional wisdom that cyber operations, however disruptive, remain categorically distinct from kinetic force. AI challenges this assumption at its foundation. That said, the assumption was already empirically fragile before the advent of generative AI. The 2007 Aurora Generator Test at Idaho National Laboratory demonstrated that a cyberattack—overriding a diesel generator's protective relays to force repeated out-of-phase connections with the grid—could physically destroy a 27-ton generator and achieving effects functionally equivalent to a kinetic strike, without a single line of traditional malware.[80] But Aurora required a bespoke setup against a single target. AI transforms this proof-of-concept into a scalable operational capability.

Autonomous agents can specifically target and poison backup systems and integrity logs over extended dwell times, modifying data blocks in ways that are statistically invisible. When recovery merely restores a compromised or corrupted state, the destruction of data is effectively permanent. This is not a theoretical concern: the operational capability for sustained, autonomous access maintenance demonstrated in Section 3.2 provides exactly the mechanism by which backup corruption at scale becomes feasible. An agent that maintains persistent access to a target's backup infrastructure for weeks, introducing subtle modifications below detection thresholds, does not need to destroy the primary systems at all; it needs only to ensure that when primary systems are restored, they are restored to a state the attacker controls. While outside of the scope of this paper, the sprawling attack surface introduced by agentic deployments within organizations further exacerbates these concerns in a strategic context. If these agents are targeted and compromised without detection, they can operate as authorized actors with human-level access across networks and systems.

The mechanism that induces an irreversible outcome is temporal asymmetry. AI increases the velocity of attack (V_a) to machine speed, while human-led recovery (V_r) remains linear and slow. As V_a exceeds V_r, the cumulative backlog of unmitigated effects creates a state of permanent degradation. Section 2.3's strategic weapons framework defines exactly this threshold: enduring effects that the target cannot reverse faster than they are imposed, sufficient to influence state-level decision-making. The V_a/V_r framework provides a quantitative basis for cross-domain

equivalence. The benchmarks in Section 3.3 provide the numerator, and infrastructure resilience metrics provide the denominator. For the first time, the comparison between a cyberattack and a kinetic strike becomes calculable rather than metaphorical.

Consider the electric grid as a theoretical scenario. The Aurora Generator Test demonstrated that a single cyberattack could physically destroy a generator by exploiting protective relay vulnerabilities. An AI-enabled campaign could extend this logic across an entire regional grid, such as simultaneously exploiting SCADA vulnerabilities at multiple substations: cycling protective relays out of phase at one facility, corrupting load-balancing algorithms at another, and manipulating frequency regulation at a third. Each fault would present differently, overwhelming operators trying to diagnose whether the cause is mechanical, environmental, or cyber-induced. Recovery from a single substation compromise typically requires days of forensic analysis, manual inspection, and controlled restart. By contrast, recovery from simultaneous, continuously evolving disruptions across a regional grid has no established timeline because the scenario has never been tested at scale.

In this scenario, an AI agent generating novel disruption sequences operates at machine speed. Human responders cannot match that pace: they are constrained by physical inspection, manual override, and replacement lead times for destroyed equipment. When V_a outpaces V_r, the grid cannot return to normal operation faster than new faults are introduced. From the perspective of the affected population, a grid held in that state for weeks is indistinguishable from a grid that has been physically destroyed. Further, the variance in replacement timeline (between 18-months to three years) for extra-high-voltage transformers means that if the attack progresses from relay manipulation to equipment destruction, the damage becomes irreversible in the most literal sense.[81]

The challenge, addressed in the next section, is that no existing risk framework is designed to integrate these measurements into policy, which hampers nation-state deterrence calculuses, resource allocation, and international norm-setting.

4.0 The Policy Landscape

The preceding sections established that AI is transforming offensive cyber capability toward the strategic threshold, and that benchmarks can quantify this transformation. Today, AI system

performance is measured by its effectiveness at a particular capability. Capability measurements will have limited utility across both civilian and military operational landscapes, unless these measurements are translated to usable or enforceable standards. As a consequence, multilateral institutions—which have historically been entrusted with the governance of global issues[82]—have struggled to establish consensus and enforceability around AI safety and security standards. In the meantime, actors continue to push the boundaries of AI-enabled operations. Because offensive capabilities inform defensive responses, policy solutions should prioritize the hardening or resilience of software systems at scale. This section explores how the United States is well positioned to spearhead this effort.

4.1 Risk Frameworks Were Not Built for Offensive Scenarios

As the nature of the threat evolves from static malware to dynamic agents, the frameworks used to assess and manage AI risk must undergo a paradigm shift. Most industry standards, while foundational, struggle to capture the adversarial and autonomous nature of AI-enabled cyber conflict and cannot comprehensively answer the question that matters most for deterrence calculus: at what point does an AI-enabled cyber capability become strategically equivalent to a kinetic one?

NIST AI Risk Management Framework (RMF). First released in January 2023 and updated in July 2024 to account for generative AI, the NIST AI RMF is the prevailing standard for managing AI risks, organized around four core functions: Govern, Map, Measure, and Manage.[83] While effective for assessing static models and general-purpose AI, it faces two significant limitations when applied to measuring offensive agentic AI. First, the RMF was designed for systems with fixed parameters and predictable deployment environments; it cannot account for agents that continuously modify their behavior and develop novel strategies during operations. Notably, these capabilities can be adapted through fine-tuning, prompt engineering, and agentic workflows without any change to underlying model weights.

Second, the AI RMF focuses on unintended failures (i.e., bias, hallucinations, and reliability) rather than intentional adversarial misuse. It lacks metrics for breakout scenarios where an agent or an attacker overseeing agentic workflows intentionally circumvents guardrails. The concept of "safety" in the RMF is oriented toward preventing harm to the user; in offensive cyber scenarios,

causing harm to the target is the objective. In December 2025, NIST released NIST IR 8596, addressing three key focus areas: securing AI system components, conducting AI-enabled cyber defense, and thwarting AI-enabled cyberattacks.[84] This represents meaningful progress, but the fundamental problem remains: the frameworks are designed for point-in-time compliance verification, not continuous assessment of adversarial AI capabilities that are improving on timescales of weeks to months.

MITRE ATLAS. MITRE ATLAS maps AI vulnerabilities similarly to how the ATT&CK framework maps cyber-centric threats.[85] However, ATLAS is primarily defensive in nature: it catalogs how to attack AI, but does not measure the strategic weight of AI-driven cyber campaigns. Like the AI RMF, ATLAS focuses almost exclusively on AI as a target rather than AI as the weapon. By prioritizing the technical vulnerabilities of the systems themselves—such as data poisoning or prompt injection—ATLAS neglects how compromise causes systemic shifts, where AI-driven automation would facilitate adversary operations at a velocity that outpaces human-centric defensive cycles. Ultimately, while ATLAS provides a necessary taxonomy for securing models, it lacks the multi-dimensional metrics required to evaluate the scalability and geopolitical intent behind AI-augmented offensive operations.

Cisco Integrated AI Security and Safety Framework. The Cisco AI Security framework advances the field with lifecycle-aware threat and risk tracking, explicitly accounting for multi-agent orchestration and model context protocols, and integrated metrics that move closer to kinetic damage assessments.[86] Yet even the Cisco framework was designed to protect organizations from AI threats (both compromises to AI and utilization of AI to compromise other systems), and an operational framework for defensive teams, not to measure offensive strategic capacity. While Cisco provides a superior tactical shield, it does not function as a strategic lens for quantifying the AI-driven overmatch potential of a nation-state's offensive capabilities.

4.2 The Benchmark-to-Policy Gap

Section 3.3 established that benchmarks can quantify the trajectory of AI-enabled offensive capability. But the relationship between benchmark performance and operational reality is more fraught than the data alone suggests. Benchmarks are inherently use-case specific, and cyber operations encompass an extraordinarily diverse range of activities across the cyber kill chain—

from reconnaissance to exploitation to sustained lateral movement—that no single evaluation comprehensively captures. The cyber-kinetic domain distinction compounds the problem: in kinetic warfare gaming, variables like range and ammunition provide natural constraints, while AI agents in cyberspace face a theoretically infinite action space with non-deterministic outcomes. Can an agent replicate a successful intrusion, or was one success an artifact of favorable conditions?

Most critically, benchmarks systematically underestimate operational capability. Lin et al. found that leading foundation models score around 50% or below on existing cybersecurity benchmarks such as Cybench, CVEBench, and BountyBench.[87] But the same researchers demonstrated that when paired with appropriate scaffolding (such as the ARTEMIS multi-agent framework), AI systems can discover vulnerabilities at rates competitive with human professionals. In a comparative study against ten cybersecurity experts on an 8,000-host university network, ARTEMIS placed second overall. The gap between benchmark performance and operational capability is itself a measurement failure: policymakers relying on benchmark scores to assess threat levels are systematically underestimating what deployed AI systems can actually do.

Emerging frameworks are beginning to address this gap. MITRE's OCCULT framework maps agent performance directly against the ATT&CK matrix, LLM use cases (i.e., knowledge assistance, autonomous operators), and reasoning power (i.e., an agent's ability to plan, observe, iterate, and generalize against evolving network defenses).[88] Dreadnode has also developed complementary practical methodologies focused on quantifying automation advantage: PentestJudge evaluates whether agents meet operational requirements including tradecraft;[89] AIRTBench measures autonomous red teaming across realistic CTF challenges;[90] and Dreadnode's action space design research examines how to structure agent environments for meaningful evaluation.[91] Together these tools provide granular visibility into how agents operate, which is a critical distinction for assessing strategic capability; however, they remain research tools, not policy instruments. The path from measurement to procurement, deterrence, and compliance decisions requires institutional investment and political will.

4.3 The International Response: Governance Without Measurement

The United Nations (UN) has emerged as the de facto primary venue for multilateral AI governance, but its approach reveals a structural mismatch with the changing threat landscape described in the preceding sections. UN discourse frames AI as a general-purpose, dual-use technology with risks that scale through integration into high-stakes systems—healthcare, public services, critical infrastructure—and can be catastrophic when combined with malign use.[92] This approach is similar to existing frameworks described in Section 4.1: it addresses how AI might cause harm through negligence or misuse, but it does not address how AI transforms the strategic weight of offensive cyber operations against those same systems.

Supranational institutions such as the United Nations have nonetheless made several advances in establishing normative policies around AI. In March 2024, the UN General Assembly's Resolution 78/265 was adopted unanimously by 193 member states, established a global baseline for “safe, secure and trustworthy AI” across its full lifecycle, though it explicitly limited its scope to the non-military domain.[93] In the cyber domain, UN member states have officially recognized threats such as ransomware attacks on hospitals and pipelines as risks to international security, which lays the groundwork for reputational enforcement through naming, attribution, and collective condemnation, as occurred following Russia's 2017 NotPetya attack on Ukrainian infrastructure.[94]

The normative value of these agreements remains important for setting baseline expectations of nation-state behavior, but it faces a limitation: reputational enforcement depends on normative consolidation, a subject matter that is currently fractured. The UN has recognized this concern as well, stressing that norms become enforceable only with broad buy-in and compliance monitoring.[95] At the Paris AI Action Summit in 2025, the United States and United Kingdom did not sign a declaration consisting of cooperative AI development principles that over 60 countries—including European Union countries, China, and India—signed.[96,97] In 2026, the United States also withheld endorsement of the International AI Safety Report, a comprehensive scientific assessment of frontier AI risks produced by a global expert panel mandated by the 2023 Bletchley Park Summit.[98,99]

The lack of consensus creates a governance vacuum with direct implications for nation-states calculus of investments and policies around AI capabilities. Norms can shape behavior through

reputational costs, but only when those norms are broadly internalized and the costs of violation are credible. As demonstrated throughout this paper, the international community has neither the measurement infrastructure to detect when AI-enabled cyber operations cross a strategic threshold, nor the enforcement mechanisms to respond when they do. Given the absence of enforceable international governance, the most effective near-term policy response would be to mandate the resilience of the systems that malicious actors could target with AI.

4.4 From Risk Aversion to Resilience by Design

Measurement tools exist (Section 3.3), while risk frameworks lag behind them (Section 4.1), and international governance organizations cannot agree on how to properly operationalize them (Section 4.4). The strategic threat described in the preceding sections cannot be addressed by regulating offensive AI capabilities directly: open-weight models capable of offensive tasks are already widely available, can be fine-tuned without oversight, and deployed on self-hosted infrastructure with no audit trail (Section 3.3.1). The viable near-term policy lever then becomes focused on hardening the target rather than controlling the weapon. In the digital era, that target is overwhelmingly civilian: the software supply chains, critical infrastructure systems, and open source codebases that governments and the global economy depend on. The Law of Armed Conflict's distinction between combatant and civilian infrastructure is a legal framework that adversaries have not historically observed in practice.

In this section, we focus on the United States as a proof-of-concept for generating momentum into multilateral or international applicability, norms development, and technological resilience. The federal government relies on open source software (OSS) for mission-critical operations, yet as adversarial AI agents capable of automated exploitation proliferate, the maintainers defending this infrastructure lack both incentives and resources. DARPA's AI Cyber Challenge (AIxCC) has demonstrated that AI can autonomously find and fix vulnerabilities in critical OSS, but current federal acquisition policies make it difficult to channel support to maintainers. Many OSS projects operate as independent contributors who cannot meet statutory requirements for federal contracting and/or prefer operational independence. Markets alone will not solve this: industry players assume someone else will cover security patches for their free software. To

actualize the promise of programs like AIxCC, a viable procurement mechanism must compensate human maintainers for verifying and integrating AI-generated patches.

To bridge this gap, a funding model built on machine-readable compliance could be considered, and could provide guidelines for federally approved, "Energy Star" or "U.S. Cyber Trust Mark"-style cybersecurity evaluations tailored for open source repositories[100]—specifically measuring their capacity for automated vulnerability detection and remediation. Just as DoD's Cybersecurity Maturity Model Certification (CMMC)[101] and Federal Risk and Authorization Management Program (FedRAMP)[102] translated NIST frameworks into strict procurement rules, policymakers should explore mechanisms to ensure that federal open source usage meets baseline resilience thresholds. However, to avoid the arduous administrative costs historically associated with these frameworks, compliance could be automated. The FedRAMP 20x pilot is already proving this model works by shifting to machine-readable evidence, which functionally automates compliance, accelerates the feedback loop, and enables continuous monitoring.

By establishing procurement guidelines rooted in machine-readable security artifacts backed by both the Cybersecurity and Infrastructure Security Agency (CISA) and NIST there is an opportunity to create a new market. Maintainers are compensated to verify AI-generated patches and produce essential chain-of-custody artifacts as commercial products. This transforms regulatory pressure into recurring revenue, securing the supply chain without the prohibitive delays of traditional federal contracting. To jumpstart this ecosystem, the federal government's Technology Modernization Fund should invest in CISA to operationalize these evaluations, partnering with the Agentic AI Foundation and its member organizations to drive adoption across leading open source AI projects.[103]

The necessity of federal mandates becomes clearer when the alternative mechanisms are examined. The private insurance market has effectively withdrawn from the risk. In August 2022, Lloyd's of London directed all syndicates to exclude losses arising from state-backed cyberattacks from standalone cyber policies, effective March 2023, citing the potential for systemic losses exceeding what the market can absorb.[104][105] The most widely adopted exclusion clause covers any nation-state cyber operation that significantly impairs a state's ability to function, precisely the class of attack this paper describes approaching feasibility. On the

enforcement side, DOJ's Civil Cyber-Fraud Initiative has recovered over $50 million through the False Claims Act[106], but it targets contractor compliance failures, not victim remediation, and the major fraud statute (18 U.S.C. § 1031) requires a minimum loss threshold of $1 million to bring charges against the United States, a threshold that leaves the vast majority of civilian cyber victims, including small municipalities, hospitals, and school districts, without a viable path to federal recourse.[107] The insurance market has priced nation-state cyber risk as uninsurable; the enforcement apparatus addresses fraud, not harm. Federal resilience mandates thus become one of the only mechanisms that remain.

The logic of this proposal extends well beyond open source software. The pipeline it would establish remains domain agnostic across AI-enabled evaluation, machine-readable compliance artifacts, procurement thresholds, and market incentives. The same approach could impose resilience baselines on any software the federal government depends on (i.e., critical infrastructure SCADA systems to state and local government networks). Investments in this approach are likely a fraction of conventional defense procurement: automated compliance pipelines cost orders of magnitude less than a single fighter jet, and they protect the civilian economic base that the military cannot function without. The cost of inaction would directly impact the strategic viability of the nation’s entire digital infrastructure.

5.0 Conclusion

The accelerated development and enablement introduced by generative AI and agentic AI systems have begun to erode historic limitations on strategic cyber capabilities, while the supporting policy and normative infrastructure meant to measure, govern, and defend against this transformation is fundamentally unprepared. Cyber effects have historically been confined to the gray zone: tactically useful under conditions of overmatch, but unable to produce decisive results at parity. The complexity of modern networks required specialized human operators to identify targets, craft bespoke payloads, and navigate compromised networks to avoid detection. AI is removing those constraints across the full offensive pipeline, from state actors embedding LLMs and agents into live operations to autonomous systems discovering zero-day vulnerabilities at scale. The barriers to strategic, scalable cyber operations are now primarily organizational, not technical or financial.

Against the backdrop of this changing landscape, existing risk frameworks were either designed for static models or for enabling defensive postures. International governance lacks both the measurement infrastructure and enforcement mechanisms to establish meaningful consensus and governance. The benchmarks that could underpin a measurement regime exist; however, the policy frameworks that should use them do not.

The path forward requires connecting measurement to mandated resilience. This paper has proposed a concrete mechanism as a proof-of-concept: machine-readable compliance standards for open source software security, modeled on the FedRAMP 20x pilot, that create procurement thresholds and market incentives for AI-assisted vulnerability remediation. The pipeline would remain domain-agnostic and extensible to any software the federal government depends on. It is not a substitute for international governance but serves as a domestic resilience baseline that protects both military readiness and the civilian infrastructure that sustains it.

Three proposals in this paper would require further research. First, the V_a/V_r framework proposed here requires empirical calibration against specific infrastructure sectors to move from theoretical construct to operational planning tool. Second, the offense-defense balance in AI-enabled cyber remains an open and consequential question; structural asymmetries currently favor offense, but defensive AI capabilities are advancing rapidly and the equilibrium may shift. Third, the systematic underestimation of operational capability by current benchmarks means that the threat trajectory presented here is likely conservative; closing the measurement gap between controlled evaluations and deployed systems is essential for credible policy.

Evidence presented in this paper suggests that convergent advancement of AI across every phase of offensive cyber operations has effectively shrunk the gray zone. The policy question that remains is whether the United States (and other nations) will build the measurement and resilience infrastructure to meet this moment, or whether it will discover the answer to that question the hard way. Resilience is not a policy preference. It is a protective force for a world in the new AI-enabled era of cyber operations.

Appendix A: Select LLM Benchmarks (Nov 2023–Feb 2026) across capability areas that can serve as proxies for offensive cyber capability areas.

AI Benchmarks: Targeting & Reconnaissance

Deep research, browsing, and multi-hop information retrieval capabilities

Benchmark	Model	Date	Score	Type
BrowseComp	OpenAI o1	2024-09	**9.9%**	Priv.
BrowseComp	GPT-4o	2024-11	**0.6%**	Priv.
BrowseComp	GPT-4o w/ browsing	2024-11	**1.9%**	Priv.
BrowseComp	GPT-4.5	2025-02	**0.9%**	Priv.
BrowseComp	OpenAI Deep Research (pass@1)	2025-02	**51.5%**	Priv.
BrowseComp	OpenAI Deep Research (best-of-N)	2025-02	**78.0%**	Priv.
BrowseComp	OpenAI o3	2025-04	**49.7%**	Priv.
BrowseComp	GPT-5 (standalone)	2025-08	**54.9%**	Priv.
BrowseComp	ChatGPT Agent (GPT-5)	2025-08	**68.9%**	Priv.
BrowseComp-Plus	Search-R1 (Qwen 32B + BM25)	2025-06	**3.9%**	OSS
BrowseComp-Plus	GPT-5 (BM25 retriever)	2025-08	**55.9%**	Priv.
BrowseComp-Plus	GPT-5 (Qwen3-Embedding-8B)	2025-08	**70.1%**	Priv.
BrowseComp-Plus	gpt-oss-20b	2025-08	**32.2%**	OSS
DeepResearchBench	Gemini 2.5 Pro Deep Research	2025-06	**48.9%**	Priv.
DeepResearchBench	OpenAI Deep Research	2025-06	**47.0%**	Priv.
DeepResearchBench	Perplexity Deep Research	2025-06	**90.2%**	Priv.
DeepResearchBench	LangChain (GPT-4.1 + Tavily)	2025-08	**6th place (no numeric score)**	Priv.
DeepResearchBench	CellCog.ai	2026-02	**54.6%**	Priv.
DeepResearchBench	Onyx Deep Research	2026-02	**54.5%**	OSS
DeepResearchBench	Qianfan-DeepResearch Pro	2026-02	**1st place Feb 3 (no numeric ...**	Priv.
FRAMES	Gemini Pro 1.5 (no retrieval)	2024-09	**40.0%**	Priv.
FRAMES	Gemini Pro 1.5 (multi-step retrieval)	2024-09	**66.0%**	Priv.
FRAMES	Gemini Flash 1.5 (no retrieval)	2024-09	**35.0%**	Priv.
FRAMES	Gemma2-27B (no retrieval)	2024-09	**25.0%**	OSS
FRAMES	Llama 3.2-3B-I (no retrieval)	2024-09	**15.0%**	OSS
GAIA	GPT-4 (plugins)	2023-11	**15.0%**	Priv.
GAIA	KGoT (GPT-4o mini)	2025-01	**34.5%**	Priv.
GAIA	OpenAI Deep Research	2025-02	**75.0%**	Priv.
GAIA	h2oGPTe (Claude 3.7 Sonnet)	2025-03	**74.0%**	Priv.
LiveDRBench	ChatGPT o3	2025-05	**1st place (no numeric score ...**	Priv.
LiveDRBench	OpenAI Deep Research	2025-05	**Below o3 (no numeric score p...**	Priv.
LiveDRBench	Perplexity	2025-05	**Mid-tier (no numeric score p...**	Priv.

AI Benchmarks: Vulnerability & Exploit Development

Code generation, vulnerability reproduction, and exploit development capabilities

Benchmark	Model	Date	Score	Type
AIxCC (DARPA)	Team Atlanta (Georgia Tech)	2024-08	**Semi-final winner (qualitati...**	OSS
Big Sleep	Gemini-based agent	2024-11	**1 zero-day found in SQLite (...**	Priv.
BigCodeBench	GPT-4o (Complete)	2024-06	**61.1%**	Priv.
BigCodeBench	GPT-4o (Instruct)	2024-06	**51.1%**	Priv.
BigCodeBench	DeepSeek-Coder-V2	2024-06	**2nd tier Elo (no numeric)**	OSS
BigCodeBench	Claude 3 Opus	2024-06	**Top-5 overall (no numeric)**	Priv.
BountyBench	Claude 3.7 Sonnet Thinking (Custom)	2025-05	**67.5%**	Priv.
BountyBench	Claude Code	2025-05	**57.5%**	Priv.
BountyBench	Claude Code	2025-05	**87.5%**	Priv.
BountyBench	Custom Agent (GPT-4.1)	2025-05	**40.0%**	Priv.
BountyBench	Custom Agent (GPT-4.1)	2025-05	**45.0%**	Priv.
BountyBench	Custom Agent (Gemini 2.5 Pro)	2025-05	**50.0%**	Priv.
BountyBench	Custom Agent (Gemini 2.5 Pro)	2025-05	**55.0%**	Priv.
BountyBench	Codex CLI (o3-high)	2025-09	**47.5%**	Priv.
BountyBench	Codex CLI (o3-high)	2025-09	**90.0%**	Priv.
BountyBench	Codex CLI (o4-mini)	2025-09	**32.5%**	Priv.
BountyBench	Codex CLI (o4-mini)	2025-09	**90.0%**	Priv.
BountyBench	Custom Agent (DeepSeek-R1)	2025-09	**37.5%**	OSS
BountyBench	Custom Agent (DeepSeek-R1)	2025-09	**25.0%**	OSS
BountyBench	Custom Agent (Qwen3 235B A22B)	2025-09	**40.0%**	OSS
BountyBench	Custom Agent (Qwen3 235B A22B)	2025-09	**45.0%**	OSS
BountyBench	Custom Agent (Llama 4 Maverick)	2025-09	**17.5%**	OSS
BountyBench	Custom Agent (Llama 4 Maverick)	2025-09	**35.0%**	OSS
BountyBench	Codex CLI o3-high (Detect)	2025-09	**12.5%**	Priv.
CyberGym	OpenHands + Claude Sonnet 4 (no thinking)	2025-06	**17.9%**	Priv.
CyberGym	OpenHands + Claude 3.7 Sonnet	2025-06	**11.9%**	Priv.
CyberGym	OpenHands + GPT-4.1	2025-06	**15.0%**	Priv.
CyberGym	SWE-Gym-32B (fine-tuned)	2025-06	**2.0%**	OSS
CyberGym	OpenHands + GPT-5 (thinking)	2025-10	**22.0%**	Priv.
CyberGym	OpenHands + GPT-5 (no thinking)	2025-10	**7.7%**	Priv.
CyberGym	Claude Sonnet 4.5 (single run)	2025-10	**28.9%**	Priv.
CyberGym	Claude Sonnet 4.5 (30 trials)	2025-10	**66.7%**	Priv.
CyberGym	GPT-5 (zero-day discovery)	2025-10	**22 confirmed zero-days from ...**	Priv.
LiveCodeBench	Gemini 3 Pro	2025-11	**91.1%**	Priv.
LiveCodeBench	Claude Opus 4.5	2025-11	**34.3%**	Priv.
LiveCodeBench	GPT-5.2	2025-12	**80.2%**	Priv.
LiveCodeBench	DeepSeek V3.2 Speciale	2025-12	**IOI Gold medalist (no Elo pu...**	OSS
MAPTA	MAPTA framework (GPT-4 + tools)	2024-06	**multi-agent pentest pipeline...**	Priv.
SWE-Bench Verified	GPT-5	2025-08	**74.9%**	Priv.
SWE-Bench Verified	Qwen3-Coder-Next (3B active)	2025-08	**70.6%**	OSS
SWE-Bench Verified	Claude Sonnet 4.5	2025-09	**77.2%**	Priv.
SWE-Bench Verified	Claude Opus 4.5	2025-11	**80.9%**	Priv.
SWE-Bench Verified	Gemini 3 Pro	2025-11	**76.2%**	Priv.
SWE-Bench Verified	GPT-5.1	2025-11	**76.3%**	Priv.
SWE-Bench Verified	GPT-5.2	2025-12	**80.0%**	Priv.
SWE-Bench Verified	DeepSeek V3.2	2025-12	**73.0%**	OSS
SWE-Bench Verified	Claude Opus 4.6	2026-02	**80.8%**	Priv.
SWE-Bench Verified	MiniMax M2.5	2026-02	**80.2%**	OSS
SWE-Bench Verified	GLM-5	2026-02	**77.8%**	OSS
SWE-Bench Verified	Kimi K2.5	2026-02	**76.8%**	OSS
XBOW	XBOW system (agentic)	2024-12	**96.3%**	Priv.

AI Benchmarks: Long-Horizon & Context Management

Sustained multi-step operations, context retention, and autonomous task completion

Benchmark	Model	Date	Score	Type
ARTEMIS	ARTEMIS (multi-agent)	2024-09	**agentic red-team pipeline (q...**	Priv.
AutoPenBench	AutoPenBench framework	2024-07	**standardize pentest eval (q...**	OSS
Context-Bench (Letta)	GPT-4.1	2025-04	**36.1%**	Priv.
Context-Bench (Letta)	Claude Opus 4.1	2025-08	**74.4%**	Priv.
Context-Bench (Letta)	GPT-5	2025-08	**70.2%**	Priv.
Context-Bench (Letta)	Claude Sonnet 4.5	2025-09	**76.5%**	Priv.
Context-Bench (Letta)	Claude Haiku 4.5	2025-10	**69.7%**	Priv.
Context-Bench (Letta)	Claude Opus 4.5	2025-11	**75.5%**	Priv.
Context-Bench (Letta)	Gemini 3 Pro	2025-11	**72.7%**	Priv.
Context-Bench (Letta)	GPT-5.2 (high)	2025-12	**77.5%**	Priv.
Context-Bench (Letta)	DeepSeek Chat (V3.2)	2025-12	**75.3%**	OSS
Context-Bench (Letta)	GLM-4.6	2025-12	**65.9%**	OSS
Context-Bench (Letta)	Claude Opus 4.6	2026-02	**82.5%**	Priv.
Incalmo	Multi-LLM orchestration	2024-07	**multi-hop attack chains (qua...**	Priv.
METR Time Horizons	Claude 3.7 Sonnet	2025-04	**12.5%**	Priv.
METR Time Horizons	Grok 4	2025-07	**high variance (no single num...**	Priv.
METR Time Horizons	GPT-5	2025-08	**28.5%**	Priv.
METR Time Horizons	GPT-5.1-Codex-Max	2025-11	**36.0%**	Priv.
METR Time Horizons	GPT-5.1-Codex-Max	2025-11	**6.7%**	Priv.
METR Time Horizons	Claude Opus 4.5	2025-12	**60.2%**	Priv.
METR Time Horizons	Claude Opus 4.5	2025-12	**5.6%**	Priv.
METR Time Horizons	Trend (all models)	2026-01	**doubling every ~4-7 months (...**	
METR Time Horizons	Gemini 3 Pro	2026-02	**evaluated Feb 2026 (score pe...**	Priv.
PentestGPT	GPT-4 (PentestGPT framework)	2024-02	**guided pentest (qualitative)**	Priv.

Endnotes

[1] U.S. Department of Defense. (2023). *2023 Department of Defense cyber strategy summary*.
[2] Lambeth, B. S. (1992). Desert Storm and its meaning: The view from Moscow (Report No. R-4164-AF). RAND Corporation. Page 23.
[3] Zetter, K. (2014). *Countdown to zero day: Stuxnet and the launch of the world's first digital weapon*. Crown.
[4] Committee on Oversight and Government Reform, U.S. House of Representatives. (2016). *The OPM data breach: How the government jeopardized our national security for more than a generation* (Majority Staff Report, 114th Congress). U.S. Government Publishing Office.
[5] Shakarian, P. (2011). The 2008 Russian cyber campaign against Georgia. *Military Review*, *91*(6), 63–68.
[6] Faesen, L., Sweijs, T., Klimburg, A., MacNamara, C., & Mazarr, M. (2020). Case study: Countering ISIS propaganda in conflict theatres. *From blurred lines to red lines: How countermeasures and norms shape hybrid conflict* (pp. 16–31). The Hague Centre for Strategic Studies.
[7] Barnes, J. E., & Kurmanaev, A. (2025, January 15). Cyberattack in Venezuela demonstrated precision of U.S. capabilities. *The New York Times*. https://www.nytimes.com/2025/01/15/us/politics/venezuela-cyberattack-us.html
[8] Przetacznik, J., & Tarpova, S. (2022). *Russia's war on Ukraine: Timeline of cyber-attacks*. European Parliamentary Research Service (EPRS).
[9] Mueller, G., Jensen, B., Valeriano, B., Maness, R., & Macias, J. (2023). *Cyber operations during the Russo-Ukrainian War: From strange patterns to alternative futures*. Center for Strategic and International Studies.
[10] Defense Science Board. (2018). *Task force on cyber as a strategic capability* (Executive summary). U.S. Department of Defense.
[11] Dykstra, J., Inglis, C., & Walcott, T. S. (2020). Differentiating kinetic and cyber weapons to improve integrated combat. *Joint Force Quarterly*, *99*(4th Quarter), 116–123.
[12] Schelling, T. C. (1966). Arms and influence. Yale University Press.
[13] Fischerkeller, M. P., & Harknett, R. J. (2017). Deterrence is not a strategy: The relevance of cyber resilience. *Journal of Strategic Studies*, *40*(3), 383–405. https://doi.org/10.1080/01402390.2017.1287413
[14] Defense Science Board, 2018.
[15] Dykstra et al, 2020.
[16] U.S. Department of Defense. (2022). *2022 national defense strategy of the United States of America: Including the 2022 Nuclear Posture Review and the 2022 Missile Defense Review*.
[17] U.S. Department of Defense. (2023). *2023 Department of Defense cyber strategy summary*.
[18] Defense Science Board. (2017). *Task force on cyber deterrence*. U.S. Department of Defense.
[19] Gartzke, E., & Lindsay, J. R. (2017). Thermonuclear cyberwar. *Journal of Cybersecurity*, *3*(1), 37–48.
[20] Borghard, E. D., & Lonergan, S. W. (2017). The logic of coercion in cyberspace. *Security Studies*, *26*(3), 452–481.
[21] Libicki, M. C. (2009). *Cyberdeterrence and cyberwar* (MG-877-AF). RAND Corporation.
[22] Dykstra et al, 2020.
[23] Maschmeyer, L. (2021). The subversive trilemma: Why cyber operations fall short of expectations. *International Security*, *46*(2), 51–90.
[24] Gartzke & Lindsay, 2017.
[25] Dykstra et al, 2020.
[26] Maschmeyer, 2021.
[27] Libicki, 2009.
[28] Fischerkeller & Harknett, 2017.
[29] Borghard & Lonergan, 2017.
[30] Maschmeyer, 2021.
[31] DARPA. (2025). AI Cyber Challenge marks pivotal inflection point for cyber defense [Press release]. https://www.darpa.mil/news/2025/aixcc-results
[32] Google Project Zero & DeepMind. (2024, October). From Naptime to Big Sleep: Using large language models to catch vulnerabilities in real-world code [Blog post]. https://projectzero.google/2024/10/from-naptime-to-big-sleep.html
[33] OpenAI. (2024, February 14). Disrupting malicious uses of AI by state-affiliated threat actors https://openai.com/index/disrupting-malicious-uses-of-ai-by-state-affiliated-threat-actors/

[34] Anthropic. (2025b, November). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. https://assets.anthropic.com/m/ec212e6566a0d47/original/Disrupting-the-first-reported-AI-orchestrated-cyber-espionage-campaign.pdf
[35] Rempfer, K. (2026, February 13). Pentagon used Anthropic's Claude during Maduro raid. *Axios*.
[36] Tau, B. (2026, February). Claude deployment via Palantir during Venezuela operation. *The Wall Street Journal*.
[37] Hazell, J. (2023). Large language models can be used to effectively scale spear phishing campaigns. *arXiv*. https://arxiv.org/abs/2305.06972
[38] Heiding, F., Lermen, S., Kao, A., Schneier, B., & Vishwanath, A. (2024). Evaluating large language models' capability to launch fully automated spear phishing campaigns: Validated on human subjects. *arXiv*. https://arxiv.org/abs/2412.00586
[39] Davies, H., McKernan, B., & Sabbagh, D. (2023, December 1). 'The Gospel': How Israel uses AI to select bombing targets in Gaza. The Guardian. https://www.theguardian.com/world/2023/dec/01/the-gospel-how-israel-uses-ai-to-select-bombing-targets
[40] Abraham, Y., Mednick, S., & Davies, H. (2023, November 30). 'A mass assassination factory': Inside Israel's calculated bombing of Gaza. *+972 Magazine*.
[41] Newman, M. (2023, July 16). Israel quietly embeds AI systems in deadly military operations. Bloomberg. https://www.bloomberg.com/news/articles/2023-07-16/israel-using-ai-systems-to-plan-deadly-military-operations
[42] Gity, R. (2025, July 20). Artificial intelligence on the battlefield in 2025. *The Jerusalem Post*. https://www.jpost.com/defense-and-tech/article-861611
[43] Google Project Zero & DeepMind, 2024.
[44] Google Cloud. (2025, July 17). Our Big Sleep agent makes a big leap. *CISO Perspectives*. https://cloud.google.com/blog/products/identity-security/cloud-ciso-perspectives-our-big-sleep-agent-makes-big-leap
[45] DARPA. (2025). AI Cyber Challenge marks pivotal inflection point for cyber defense [Press release]. https://www.darpa.mil/news/2025/aixcc-results
[46] Carlini, N., Lucas, K., Asher, E. B., Cheng, N., Lakhani, H., Forsythe, D., & Guru, K. (2026, February 5). *Evaluating and mitigating the growing risk of LLM-discovered 0-days*. Anthropic. https://red.anthropic.com/2026/zero-days/
[47] Google Threat Intelligence Group. (2025, November 5). GTIG AI threat tracker: Advances in threat actor usage of AI tools. *Google Cloud Blog*. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
[48] Akil, M. A., et al. (2025). LLMalMorph: On the feasibility of generating variant malware using large-language-models. *arXiv*. https://arxiv.org/abs/2507.09411
[49] U.S. Department of Defense. (2025). Annual report to Congress: Military and security developments involving the People's Republic of China 2025, 57.
[50] Cisco. (2026, February 19). State of AI security 2026 report. https://www.cisco.com/site/us/en/products/security/state-of-ai-security.html
[51] Splunk Threat Research Team. (2025, September 25). From prompt to payload: LAMEHUG's LLM-driven cyber intrusion analysis. https://www.splunk.com/en_us/blog/security/lamehug-ai-driven-malware-llm-cyber-intrusion-analysis.html
[52] Google Threat Intelligence Group. (2025, November 5). GTIG AI threat tracker: Advances in threat actor usage of AI tools. *Google Cloud Blog*. https://cloud.google.com/blog/topics/threat-intelligence/threat-actor-usage-of-ai-tools
[53] Google Threat Intelligence Group, 2025.
[54] Ilascu, I. (2025, November 5). Hackers are already using AI-enabled malware, Google says. *Axios*.
[55] Google Threat Intelligence Group, 2025.
[56] Google Threat Intelligence Group, 2025.
[57] Ilascu, 2025.
[58] Ukraine State Service for Special Communications and Information Protection. (2025). *H1 2025 threat report*.
[59] Kovacs, E. (2025, October 9). From phishing to malware: AI becomes Russia's new cyber weapon in war on Ukraine. *The Hacker News*.
[60] Anthropic. (2025a, August 27). *Threat intelligence report: August 2025*. https://www-cdn.anthropic.com/b2a76c6f6992465c09a6f2fce282f6c0cea8c200.pdf
[61] Anthropic, 2025a.
[62] +972 Magazine. (2025, March 6). Israel developing ChatGPT-like tool that weaponizes surveillance of Palestinians.
[63] Biesecker, C., Mednick, S., & Burke, J. (2025, February 18). As Israel uses US-made AI models in war, concerns arise about tech's role in who lives and who dies. *Associated Press*.

[64] +972 Magazine. (2025)
[65] Newman, 2023.
[66] Iran International. (2025, July 21). Israel-backed cyberattacks cripple IRGC finances, burn $90 million in crypto—WSJ.
[67] Kapko, M. (2025, June 17). Iran's Bank Sepah disrupted by cyberattack claimed by pro-Israel hacktivist group. *CyberScoop*. https://cyberscoop.com/iran-bank-sepah-cyberattack/
[68] Gity, R. (2025, July 20). Artificial intelligence on the battlefield in 2025. The Jerusalem Post. https://www.jpost.com/defense-and-tech/article-861611
[69] Gity (2025).
[70] Picus Security. (2025, November 4). Predatory Sparrow: Inside the cyber warfare targeting Iran's critical infrastructure. https://www.picussecurity.com/resource/blog/predatory-sparrow-inside-the-cyber-warfare-targeting-irans-critical-infrastructure
[71] NPR. (2021, October 27). A cyberattack paralyzed every gas station in Iran. https://www.npr.org/2021/10/27/1049566231/irans-president-says-cyberattack-was-meant-to-create-disorder-at-gas-pumps
[72] SCADAfence. (n.d.). Industrial cyber attack on Iranian steel companies explained. *Claroty Blog*. https://blog.scadafence.com/the-iran-steel-industry-cyber-attack-explained
[73]Malwarebytes. (2022, July 14). Predatory Sparrow massively disrupts steel factories while keeping workers safe. *Malwarebytes Blog*. https://www.malwarebytes.com/blog/news/2022/07/predatory-sparrow-massively-disrupts-steel-factories-while-keeping-workers-safe
[74] Kapko (2025).
[75] TechCrunch. (2025, June 17). Pro-Israel hacktivist group claims responsibility for alleged Iranian bank hack. https://techcrunch.com/2025/06/17/pro-israel-hacktivist-group-claims-responsibility-for-alleged-iranian-bank-hack/
[76] TRM Labs. (2025, June). Inside the Nobitex breach: What the leaked source code reveals about Iran's crypto infrastructure. https://www.trmlabs.com/resources/blog/inside-the-nobitex-breach-what-the-leaked-source-code-reveals-about-irans-crypto-infrastructure
[77] Wall Street Journal. (2025, June 29). How Israel-aligned hackers hobbled Iran's financial system. https://www.wsj.com/world/middle-east/how-israel-aligned-hackers-hobbled-irans-financial-system-fb1b0376
[78] METR. (2026, January 29). Time Horizon 1.1 [Blog post]. https://metr.org/blog/2026-1-29-time-horizon-1-1/
[79] Gartzke, E. (2013). The myth of cyberwar: Bringing war in cyberspace back down to earth. *International Security*, *38*(2), 41–73.
[80] Zetter, K. (2014). *Countdown to zero day: Stuxnet and the launch of the world's first digital weapon*. Crown.
[81] Trabish, H. K. (2025, February 12). Transformer supply bottleneck threatens power system stability as load grows. Utility Dive. https://www.utilitydive.com/news/electric-transformer-shortage-nrel-niac/738947
[82] United Nations. (n.d.). *Multilateral system*. https://www.un.org/en/global-issues/multilateral-system
[83] National Institute of Standards and Technology. (2024). *Artificial intelligence risk management framework: Generative artificial intelligence profile* (NIST AI 600-1). https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[84] Megas, K., Cuthill, B., Snyder, J. N., Patrick, B., Khemani, I., Dotter, M., Garris, M., Zarei, M., & Schiro, N. (2025). *Cybersecurity AI profile: NIST community profile* (NIST IR 8596). National Institute of Standards and Technology.
[85] MITRE. (n.d.). *MITRE ATT&CK®*. Retrieved December 31, 2025, from https://attack.mitre.org/
[86] Chang, A., Saade, T., Mendapara, S., Swanda, A., & Garg, A. (2025). Cisco integrated AI security and safety framework report. *arXiv*. https://arxiv.org/abs/2512.12921
[87] Lin, J. W., Jones, E. K., Jasper, D. J., Ho, E. J. S., Wu, A., Yang, A. T., & Ho, D. E. (2025). Comparing AI agents to cybersecurity professionals in real-world penetration testing. *arXiv*. https://arxiv.org/abs/2512.09882
[88] Kouremetis, M., et al. (2025). OCCULT: Evaluating large language models for offensive cyber operation capabilities. *arXiv*. https://arxiv.org/abs/2502.15797
[89] Caldwell, S., Harley, M., Kouremetis, M., Abruzzo, V., & Pearce, W. (2025). PentestJudge: Judging agent behavior against operational requirements. *arXiv*. https://arxiv.org/abs/2508.02921
[90] Dawson, A., Mulla, R., Landers, N., & Caldwell, S. (2025). AIRTBench: Measuring autonomous AI red teaming capabilities in language models. *arXiv*. https://arxiv.org/abs/2506.14682
[91] Dreadnode. (2025). Evals: The foundation for autonomous offensive security [Blog post]. https://dreadnode.io/blog/evals-the-foundation-for-autonomous-offensive-security
[92] United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*. United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf

[93] United Nations General Assembly. (2024). *Seizing the opportunities of safe, secure and trustworthy artificial intelligence systems for sustainable development* (Resolution 78/265). United Nations. https://docs.un.org/en/A/res/78/265
[94] Cybersecurity and Infrastructure Security Agency. (2017, July 1). *Petya ransomware* (Alert TA17-181A). U.S. Department of Homeland Security. https://www.cisa.gov/news-events/alerts/2017/07/01/petya-ransomware
[95] United Nations Secretary-General's High-Level Advisory Body on Artificial Intelligence. (2024). *Governing AI for humanity: Final report*. United Nations. https://www.un.org/sites/un2.un.org/files/governing_ai_for_humanity_final_report_en.pdf
[96] Milmo, D., & Courea, E. (2025, February 11). US and UK refuse to sign Paris summit declaration on "inclusive" AI. *The Guardian*. https://www.theguardian.com/technology/2025/feb/11/us-uk-paris-ai-summit-artificial-intelligence-declaration
[97] Tasioulas, J. (2025, February 14). Expert comment: Paris AI summit misses opportunity for global AI governance. *University of Oxford*. https://www.ox.ac.uk/news/2025-02-14-expert-comment-paris-ai-summit-misses-opportunity-global-ai-governance
[98] Bengio, Y., Clare, S., Prunkl, C., Murray, M., Andriushchenko, M., Bucknall, B., Bommasani, R., Casper, S., Davidson, T., Douglas, R., Duvenaud, D., Fox, P., Gohar, U., Hadshar, R., Ho, A., Hu, T., Jones, C., Kapoor, S., Kasirzadeh, A., . . . Mindermann, S. (2026). *International AI Safety Report 2026* (DSIT 2026/001). Department for Science, Innovation and Technology. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
[99] UK Government. (2023, November 1). *The Bletchley Declaration by countries attending the AI Safety Summit, 1–2 November 2023*. GOV.UK. https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023
[100] Federal Communications Commission. (2025). U.S. Cyber Trust Mark. https://www.fcc.gov/CyberTrustMark
[101] Office of the Federal Register. (2024, October 15). Cybersecurity Maturity Model Certification (CMMC) Program. Federal Register, 89(199), 83092-83237. https://www.govinfo.gov/app/details/FR-2024-10-15/2024-22905
[102] U.S. General Services Administration. (2023). FedRAMP cloud service provider authorization playbook. https://www.fedramp.gov/
[103] Agentic AI Foundation (n.d.). Agentic AI Foundation. https://www.agenticai.foundation/
[104] Lloyd's. (2022, August 16). State backed cyber-attack exclusions (Market Bulletin Y5381). https://assets.lloyds.com/media/35926dc8-c885-497b-aed8-6d2f87c1415d/Y5381%20Market%20Bulletin%20-%20Cyber-attack%20exclusions.pdf
[105] Lloyd's. (2024, May 14). *State-backed cyber-attack wordings* (Market Bulletin Ref: Y5433). https://lmalloyds.com/wp-content/uploads/2025/06/Y5433.pdf
[106] Bueker, J. P., Hallward-Driemeier, D., Hoey, L. G., Kossak, A. D., Lampert, M. B., Levy, J. S., & O'Connor, A. (2026, January 20). False Claims Act insights: Key takeaways from DOJ's fiscal year 2025 cases & recoveries. Ropes & Gray. https://www.ropesgray.com/en/insights/alerts/2026/01/false-claims-act-insights-key-takeaways-from-dojs-fiscal-year-2025-cases-recoveries
[107] Major Fraud Act of 1988, 18 U.S.C. § 1031 (2018).

De-Risking Defense Innovation at the Earliest Stages: The Strategic Role of Entrepreneurial Fellowships and Early Non-Dilutive Grant Funding

Elizabeth Kennedy and Lauren Emmi

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Bios

Elizabeth Kennedy is the Director of Government Relations at Activate, a national nonprofit enabling nearly 250 hard-technology startups raising more than $5 billion total. Prior to Activate, she was Head of Corporate Development at DARPA-funded company Portal Biotechnologies, and Vice President of Business Development and Strategy at the Massachusetts Life Sciences Center, working to deploy $1 billion in state funding to support the life sciences sector. Previously, she was Chief of Staff at Landmark Bio, a Harvard-MIT-affiliated joint-venture; Special Projects Manager at Harvard for the Landmark Bio initiative under Dr. Alan Garber; and a fellow in the Kennedy School's Emerging Technology Policy program.

Lauren Emmi is a first-year MBA student at Harvard Business School and a Major in the U.S. Army Reserves. Her background includes over thirteen years as an active-duty Army intelligence officer, supporting joint intelligence and special operations across five deployments. Most recently, she was selected for and completed the U.S. Army's Congressional Fellowship, serving as Legislative Aide to the Vice Chief of Staff of the Army and as Legislative Liaison for the Army's intelligence and network portfolios, where she developed congressional engagement strategies supporting Army modernization priorities. Prior to HBS, she interned at Moonshots Capital, a veteran-founded venture capital firm investing in dual-use technologies.

Abstract

Despite billions in sustained U.S. investment in defense-related basic and applied research, many high-potential technologies fail to transition from the laboratory to operational capability, resulting in capital inefficiency, delayed warfighter access, and erosion of U.S. technological advantage. This gap is most pronounced at the earliest technology readiness levels (TRLs), where technical uncertainty, unclear demand signals, and limited access to capital stall progress. Traditional defense acquisition and grant mechanisms are optimized for integrating mature technologies against well-defined requirements, rather than high-risk early technologies, leaving early-stage innovations without viable pathways to maturation and adoption.

This paper examines entrepreneurial fellowship models, defined as structured, multi-year programs that provide early-stage, non-dilutive funding, as critical infrastructure for defense and other critical innovation. These models provide intensive mentorship through company formation, salary support enabling full-time founder commitment, and early connections to investors and government customers. By sustaining founders with modest grant funding through the highest-risk period of financing and technology validation—before conventional acquisition or venture capital funding are viable—entrepreneurial fellowships prevent loss of critically important technological innovations before they reach acquisition relevance. Beyond addressing this core failure in the innovation pipeline, they enable earlier alignment with military requirements while achieving public capital efficiency.

Drawing on evidence from defense-adjacent fellowship programs and comparable innovation initiatives, this paper argues that early-stage grant-based fellowships outperform other funding mechanisms in three key ways: (1) accelerating transition from proof-of-concept to defensible prototype, (2) reducing downstream acquisition risk by enabling earlier validation of operational use cases and requirements, and (3) crowding in private capital at later stages without prematurely forcing commercial scaling or misaligned market entry. Importantly, these programs complement—not replace—existing acquisition pathways by feeding them with better-defined, more advanced technologies. These effects are particularly pronounced in strategically critical technology domains, including position, navigation, timing (PNT) capabilities, advanced manufacturing, biomanufacturing, and advanced materials, where fellowship-supported

companies have delivered novel technologies that surpass existing solutions in both capability and cost.

In an era in which future conflicts will be decided by technological superiority, the nation that validates and fields critical technologies first accrues decisive strategic advantage. This paper concludes with policy recommendations for the Department of Defense (DoD) and interagency partners to expand and institutionalize early-stage entrepreneurial fellowship funding as a strategic tool for strengthening the defense industrial base, shortening time-to-capability, and preserving U.S. technological advantage amid intensifying global competition.

I. Introduction: The Early-Stage Defense Innovation Gap

The U.S. government invests more in defense-related research and development than any other nation.[1] In FY26, Congress appropriated nearly $146 billion for the Department of Defense (DoD) Research, Development, Test, and Evaluation (RDT&E) budget.[2] RDT&E spending as a portion of the overall topline defense budget has grown from 11% in 2012, to now approximately 15% in 2026.[3] Even as nominal RDT&E spending has grown, the U.S. share of global R&D has fallen sharply, and analysts increasingly question whether existing investments are translating into preserved technological overmatch.[4] These doubts are reinforced by the relative stagnation of the DoD procurement budget over the same time period, instead of a corresponding increase, and the increased time between program start and expected initial operational capability. From 2023 to 2024, this "cycle time" increased from 124 months, to 142 months for the DoD's Major Defense Acquisition Programs (MDAPs).[5] While difficult to track the true transition rate from R&D to achieving program of record status, it is widely-accepted that a large share of defense-funded research fails to transition into programs of record, and a significant share that do face years of delays before reaching operational use.[6]

This persistent "early-stage defense innovation gap" is not merely an administrative inefficiency; it represents a structural failure in the pre-acquisition phase of the defense innovation pipeline, where promising technologies fail before they ever reach requirements alignment, program sponsorship, or viable routes to adoption.[7] As strategic competition intensifies, this gap constitutes a material vulnerability: public capital is expended without proportional operational return, while competitors shorten the timeline from discovery to deployment and gain learning advantages through earlier field experimentation.[8]

This early-stage gap is most acute at the lower technology readiness levels (TRLs 2–4), before formal requirements exist and well before programs of record are established.[9] At this stage, innovators face technical uncertainty, unclear demand signals, and limited access to capital. Traditional defense acquisition mechanisms are designed to integrate mature technologies against defined requirements, not to support high-risk early technologies whose operational relevance is still being discovered. Meanwhile, private venture capital is often ill-suited to

finance defense-relevant technologies with long development cycles, regulatory complexity, and concentrated government customers—particularly in capital-intensive domains such as advanced materials, biomanufacturing, quantum and information systems, microelectronics, and position, navigation, and timing (PNT) capabilities.[10]

As a result, many strategically critical technologies fall into a structural "valley of death," where they are too early for procurement, too uncertain for private capital, and too capital-intensive for founders to sustain without institutional support. The consequences of this failure are both economic and strategic. First, the defense innovation system exhibits pronounced capital inefficiency. Large amounts of public research funding produce technical knowledge and early prototypes that never mature into deployable capabilities, resulting in structurally low returns on public R&D investment.[11] This inefficiency is not anomalous but systemic, reflecting long-standing misalignments between early-stage research incentives, acquisition bureaucracies optimized for mature systems, and industrial stakeholders whose business models depend on program-of-record rather than pre-acquisition experimentation.[12] Second, failure to transition early innovation erodes technological advantage: historically, U.S. military effectiveness has depended not only on scale, but on the ability to integrate novel technologies into operational systems faster than U.S. adversaries.[13] When early-stage innovations stall, the country forfeits the strategic benefits of time, iteration, and early operational learning—advantages that increasingly determine outcomes in technologically mediated competition.[14]

This paper advances a central thesis that the dedicated support of early-stage entrepreneurial fellowship programs (multi-year funding for scientists and engineers commercializing their research), paired with other non-dilutive funding sources, can often function as this missing infrastructure in the U.S. defense innovation pipeline. Properly designed, these programs de-risk the earliest stages of technology maturation and company formation by providing needed founder-capital stabilization through salary and other funding support, connections to relevant DoD funding agencies and mentors/advisors, and early exposure to DoD problem sets—before conventional acquisition mechanisms or venture capital are viable. In doing so, entrepreneurial fellowships convert fragile, early-stage research into acquisition-relevant capability pipelines, restoring public capital efficiency, accelerating time-to-capability, and preserving U.S.

technological advantage in critical technology domains. This analysis draws on primary-source interviews, program participation data, and published research on innovation financing and defense acquisition.

This paper references both historical and contemporary sources spanning multiple administrative periods. For consistency, "Department of Defense (DoD)" is used throughout, except when referencing the formal title of a specific report or official position where "Department of War (DoW)" is used. Additionally, this paper uses TRLs to describe technological maturity, rather than DoD RDT&E budget classifications, to provide a consistent framework across agencies and funding mechanisms.

II. Why Existing Defense Funding and Acquisition Mechanisms Fall Short at Early Stages

A. Limits of Traditional Research Grants

Federal research grants—particularly those structured around discrete technical milestones—have long served as a cornerstone of U.S. defense innovation policy. Programs such as the Small Business Innovation Research (SBIR) and Small Business Technology Transfer (STTR) programs are explicitly designed to stimulate early-stage technological development and are critical to expanding the industrial base.[15]

However, these awards are fundamentally project-centric, and funding is tied to specific technical deliverables, bounded timelines, and narrowly defined research objectives.[16] The organizational entity receiving the grant is often treated as a vehicle for executing a project rather than as a firm requiring capitalization, institutional development, and strategic positioning within acquisition pathways. This model is well-suited to project delivery but less aligned with building durable, operationally capable, cutting-edge technology companies. As defense acquisition increasingly depends on venture-backed startups, the project-centric nature of early-stage funding may limit the development of companies capable of sustaining long-term engagement with the DoD.[17] Without a viable pathway for early-stage startups to survive the pre-acquisition

phase, the DoD forfeits access to the most innovative segment of the technology base precisely when intervention would be cheapest and most consequential.

A second limitation concerns incentives surrounding commercialization. SBIR and related research programs formally encourage commercial transition, yet award structures primarily reward technical performance against milestone criteria rather than market or operational adoption.[18] Research grants often provide no salary stability for founders beyond the duration of the award and may restrict the allocation of funds toward business development and/or customer integration.[19] As a result, founders must frequently secure additional funding sources to sustain operations between phases, diverting attention from technical iteration and defense customer engagement.[20]

Third, traditional research grants frequently lack continuity across funding phases. Although SBIR programs include sequential phases (i.e. Phase I, II, and III), transition rates vary widely, and progression is neither automatic nor structurally guaranteed.[21] Numerous U.S. Government Accountability Office (GAO) reports have documented challenges in moving technologies from early research funding into sustained acquisition programs.[22] The discontinuities between research, prototyping, and procurement create financing gaps that are especially acute for early-stage firms with limited administrative capacity. In practice, companies may experience extended intervals between awards or confront uncertainty regarding follow-on funding, undermining organizational stability during critical technical development stages.

These structural characteristics reveal a frequent misalignment between grant-based research instruments and the needs of early-stage defense-relevant companies operating at pre-prototype stages. If modernization depends on translating laboratory breakthroughs into deployable systems, then founder-level stability and organizational continuity become strategically relevant variables. Addressing this gap does not require replacing research grants but complementing them with mechanisms that stabilize companies—not merely projects—during the highest-risk phase of technological maturation.

B. Acquisition System Constraints

The U.S. defense acquisition system is structurally optimized for integrating mature technologies against well-defined operational requirements, rather than for supporting high-risk innovations at the earliest stages of technical development. This problem is explicitly acknowledged in the 2022 National Defense Strategy (NDS), stating that the current system is, "too slow and too focused on acquiring systems not designed to address the most critical challenges".[23] Formal acquisition processes, whether through MDAP or accelerated procurement authorities, presuppose stable performance parameters, validated concepts of operation, and programmatic sponsorship within the requirements and budgeting process. This design creates a fundamental mismatch for technologies at lower technology readiness levels (TRLs 2–4), where technical feasibility, operational relevance, and integration approaches are still uncertain. As a result, the acquisition system is largely inaccessible to critical early-stage innovations, not because of managerial failure, but because it is institutionally configured to manage risk at scale rather than to absorb risk at early-stage company formation.[24]

Risk aversion is embedded in the Department's acquisition governance structures, which are designed to minimize cost overruns, schedule delays, and political exposure. Extensive documentation requirements, milestone reviews, and compliance obligations serve important oversight functions but impose disproportionate burdens on early technologies and small teams. Early-stage companies lack the historical performance data, test histories, contracting expertise, programmatic backing, security clearances, compliance infrastructure, and cash flow to navigate the acquisition process, which was designed around incumbent defense contractors with dedicated government relations teams and long-standing institutional relationships.[25]

As a result, the system is biased toward large, existing vendors rather than toward novel technologies.[26] Even when these early technologies are strategically promising and startup companies engage through pilot programs or prototype contracts, the absence of clear transition pathways into programs of record frequently leads to "pilot purgatory," where technical success does not translate into sustained procurement or operational adoption. This dynamic discourages early-stage entrants and narrows the effective supplier base, reinforcing a defense industrial structure optimized for scale rather than technological dominance.

C. Capital Market Mismatch

Venture capital has played an increasingly prominent role in defense-adjacent technology sectors over the past decade, particularly in software and artificial intelligence.[27] However, its incentive structures remain poorly aligned with the earliest stages of defense-relevant technological development—particularly in hard technology areas that include quantum, advanced materials, energy systems, and advanced manufacturing technologies.

Unlike software startups, which can iterate rapidly at relatively low marginal cost, hard technology companies must absorb expenses related to materials, tooling, fabrication, and testing infrastructure.[28] The capital intensity of early-stage hard technology development often necessitates larger initial investments prior to meaningful commercial validation. Moreover, many defense-relevant technologies may require specialized testing facilities or compliance requirements that further increase costs and extend development timelines.[29]

Additionally, venture capital funds are typically structured around fixed time horizons, with capital deployment and exit expectations to deliver returns within a five-year fund lifecycle.[30] Within this model, investments that require prolonged technical iteration and capital-intensive prototyping are inherently less attractive than software ventures capable of rapid scaling. Defense-relevant technologies at TRL 2–4 frequently require extended technical milestone cycles, laboratory refinement, and pilot-scale validation prior to revenue generation.

Dual-use ventures introduce additional complexity. Commercial venture investors often hesitate to back technologies perceived as dependent on government procurement cycles and/or geopolitical volatility.[31] At the same time, defense acquisition pathways frequently assume that companies possess sufficient private capitalization to absorb technical risk prior to award.

The existence of this capital market mismatch does not imply failure of venture capital markets; rather, it reflects the rational alignment of investment incentives with fund structure. However, when defense modernization depends on translating early-stage scientific advances into

deployable capabilities, the absence of stable founder-level support during pre-venture and pre-prototype phases becomes strategically consequential. Mechanisms that stabilize founders and absorb early technical risk—without imposing premature scaling expectations—may therefore function as complementary infrastructure to both federal research grants and private venture investment.

D. Strategic Competition and the Imperative for Speed

A defining feature of contemporary defense innovation is that much of the foundational research underpinning future military capabilities is conducted in the open. University laboratories, Federally Funded Research and Development Centers (FFRDCs), and pre-commercial startups publish findings in peer-reviewed journals, which are presented at academic conferences and disseminated through numerous websites that U.S. adversaries can easily access. Today, strategic advantage no longer flows primarily to the nation that discovers first, but to the nation that commercializes, scales, and fields first.[32] Nations that shorten this timeline gain the advantage.[33]

However, while the American system excels at generating foundational research, it remains slow to convert that research into operational capability.[34] Even alternative acquisition mechanisms designed to move technologies ahead faster, such as Other Transaction Authorities (OTAs), struggle to transition successful prototypes into sustained production.[35] By contrast, China has accelerated its ability to scale technologies in critical areas such as quantum, batteries, and autonomous systems, absorbing early technical and commercialization risk and allowing the country to bend emerging technologies toward national-priorities, including defense.[36] While the U.S. may not want to endorse China's state-led model, it should look to improve its ability to move first, from lab to field, to reduce its strategic vulnerabilities.

III. Defining Entrepreneurial Fellowship Models

A. What Is an Entrepreneurial Fellowship?

An entrepreneurial fellowship, as used in this analysis, refers to a structured, multi-year program designed to support company formation and prototype development. Unlike short-duration accelerators or project-specific research grants, entrepreneurial fellowships are organized around the stabilization and maturation of a company rather than the completion of a single technical deliverable. The central unit of support is the founder and the early company, not an isolated research milestone.

While there is some variation between entrepreneurial fellowships, these programs typically are geared to support first-time technical founders (i.e., scientists and engineers) with demonstrated scientific or engineering expertise—often requiring a bachelor's degree and multiple years of post-baccalaureate research or technology development experience. Participants are expected to be leading the commercial development of a hardware-based innovation and be able to commit full time to advancing technologies that remain in development rather than market-ready.

These programs provide early-stage, non-dilutive funding targeted at technologies that typically enter the program between TRLs 2–3, where technologies may have demonstrated conceptual feasibility but lack prototype validation, and typically exit between TRLs 4–6 with their first validated product.[37] Unlike venture capital financing, which is typically structured around expectations of rapid revenue growth and exit timelines, entrepreneurial fellowships are not predicated on short-term liquidity events.[38] Instead, they enable iterative validation without requiring premature scaling toward commercial markets. By stabilizing founders financially during this high-risk period, fellowships aim to reduce early-stage fragility without imposing premature equity financing or aggressive scaling expectations.

A defining feature of entrepreneurial fellowships is their focus on company formation and technology development as interconnected processes, rather than treating commercialization as a downstream step separate from laboratory research. These programs recognize that building a company—forming a team, engaging customers, and shaping intellectual property strategy—occurs alongside technical progress.[39] In defense-relevant fields, this is especially important, as founders must refine their technologies while also learning to navigate acquisition processes, compliance requirements, and federal contracting.

Entrepreneurial fellowships therefore act as an intermediate mechanism between research and venture-backed scale up, complementing federal grants and private capital by addressing a key structural gap: the instability faced by founders at pre-prototype stages. By providing multi-year support during this high-risk phase, fellowships increase the likelihood that early scientific advances develop into durable companies capable of sustained engagement with defense acquisition systems.

B. Core Components That Drive Effectiveness

If entrepreneurial fellowships are to function as structural complements to traditional research grants and venture capital, their effectiveness depends on several core components. These components are not incidental features of program design, but mechanisms intended to address specific vulnerabilities in early-stage, defense-relevant technology development. Four components appear central: early-stage funding including founder salary stipend, intensive mentorship and community support, early connections to investors and government funders, and structured curriculum and milestones.

Early-Stage, Non-Dilutive Funding, Including Founder Salary Stipend and Other Support

The first component is non-dilutive funding, which includes a fully funded founder salary stipend as well as funding for other expenses.[40] Additionally, non-dilutive funding for early-stage research and development to advance prototyping is also often provided.[41] Such funding enables prototype development and iterative testing without imposing immediate equity dilution or revenue pressure. In defense-relevant domains—where hardware iteration, materials testing, and compliance can be capital-intensive—this funding functions as bridge capital, stabilizing early companies prior to entry into more formal acquisition channels. In total, this funding often amounts to around $350,000 over the multi-year program, such as the National Science Foundation (NSF)'s entrepreneurial fellowship program.[42]

Intensive Mentorship and Community Support

A second component involves intensive mentorship and community support for early-stage technical founders.[43] Structured mentorship and community support introduce accountability and informed iteration without imposing venture-style scaling pressures.[44] As numerous founders observed, participation in an entrepreneurial fellowship provided mentorship and peer support that helped translate technical work into compelling proposals for DoD funding opportunities, supporting founders in developing "pitch narratives" and proposal strategies tailored to DoD audiences for Army, AFWERX, and more.[45]

Early Exposure to Government Funders and Relevant Investors

Fellowship models also provide early exposure to both government funders and investors. Engagement with defense stakeholders—such as program managers, research offices, and innovation units—can clarify operational needs and inform technical roadmaps before formal procurement begins, reducing informational gaps and potentially shortening the path to funding once proposals are submitted. Early interaction with investors likewise helps align capital strategy with the longer timelines typical of defense-relevant technologies in particular, reducing pressure for premature scaling or commercial pivoting. Together, these connections integrate company formation into broader innovation and acquisition pipelines rather than isolating technical development from downstream capital and contracting pathways. Participants noted that the program facilitated high-level government introductions that would otherwise have been difficult to obtain with offices such as the former National Security Innovation Capital (NSIC) under the Defense Innovation Unit (DIU).[46] Other participants noted how the entrepreneurial fellowship "bridged the gap between academic-stage innovation" and venture capital investment.[47]

Structured Curriculum and Milestones

Structured curricula and milestone-driven progress are central features of entrepreneurial fellowship models. Participants noted that the program's structure required founders to develop technical roadmaps with defined deliverables tied to real customer needs.[48] The programs often

include workshops, virtual trainings, and regular meetings with fellowship staff members that provide feedback and support to teams.

In aggregate, these mechanisms suggest that entrepreneurial fellowships function not as accelerators in the conventional sense, but as stabilization infrastructure during the most vulnerable phase of defense-relevant technological development.

C. How Entrepreneurial Fellowships Differ from Existing Innovation Mechanisms

Entrepreneurial fellowships operate within a broader patchwork of federal research programs, private accelerators, and university incubators. As NSF Assistant Director for Technology, Innovation, and Partnerships (TIP) Erwin Gianchandani has noted of their program, "Entrepreneurial fellowships offer another pathway for researchers to transition promising ideas and technologies from the lab to society."[49] Thus, while each of these mechanisms play a distinct role in defense innovation, fellowship models differ in structural orientation and institutional function. Understanding these distinctions clarifies why fellowships address a specific gap in early-stage defense-relevant technology development.

SBIR/STTR: Project-Centric Research Funding

The SBIR and STTR programs are among the most important federal mechanisms for stimulating early-stage technical development within small businesses. As previously discussed, these programs provide non-dilutive funding through phased awards and are explicitly intended to promote commercialization.[50] However, as mentioned, SBIR/STTR awards are fundamentally project-centric and while awards may provide critical early capital, they do not typically address founder income stability, organizational development, or sustained institutional capacity between awards.[51] Entrepreneurial fellowships differ in that they are company-centric: they aim to stabilize the venture and founder across multiple stages of technical maturation, irrespective of the outcome of a single grant application. In this sense, fellowships may complement SBIR by providing necessary support to pursue and sustain federal awards and DoD contracts.

Accelerators: Short-Duration, Scale-Oriented Programs

Private-sector accelerators are generally structured as short-duration programs (often three to six months) focused on rapid customer validation, fundraising preparation, and growth acceleration.[52] These models are typically optimized for software ventures capable of rapid iteration and market entry. By contrast, fellowship models are often multi-year, fully funded programs designed to extend technical runways during pre-prototype stages. They provide non-dilutive salary stipend support and structured technical validation rather than prioritizing near-term revenue growth. For defense-relevant hardware and manufacturing technologies, where iteration cycles are longer and commercialization timelines less predictable, the accelerator model may inadequately address early-stage fragility. Fellowships therefore appear to serve a distinct function focused on stabilization and longer-term acceleration.

Traditional Incubators: Infrastructure Without Sustained Capitalization

University-affiliated incubators and innovation hubs provide valuable physical infrastructure, networking opportunities, and advisory support.[53] However, many incubators do not provide sustained, non-dilutive funding or founder salary support. Their primary contribution lies in co-location and ecosystem integration rather than direct financial stabilization. While incubators may facilitate company formation, they do not necessarily address the capital gap experienced by founders transitioning from academic research to prototype development. By directly funding both R&D and founder salary through a multi-year program, fellowships seek to reduce the fragility that often undermines laboratory spinouts prior to defense engagement.

Additionally, many companies require access to specialized research facilities and equipment, which these incubators often do not have. Entrepreneurial fellowship programs have provided access to these facilities and equipment as a part of their programs.[54]

IV. How Entrepreneurial Fellowships De-Risk Defense Innovation

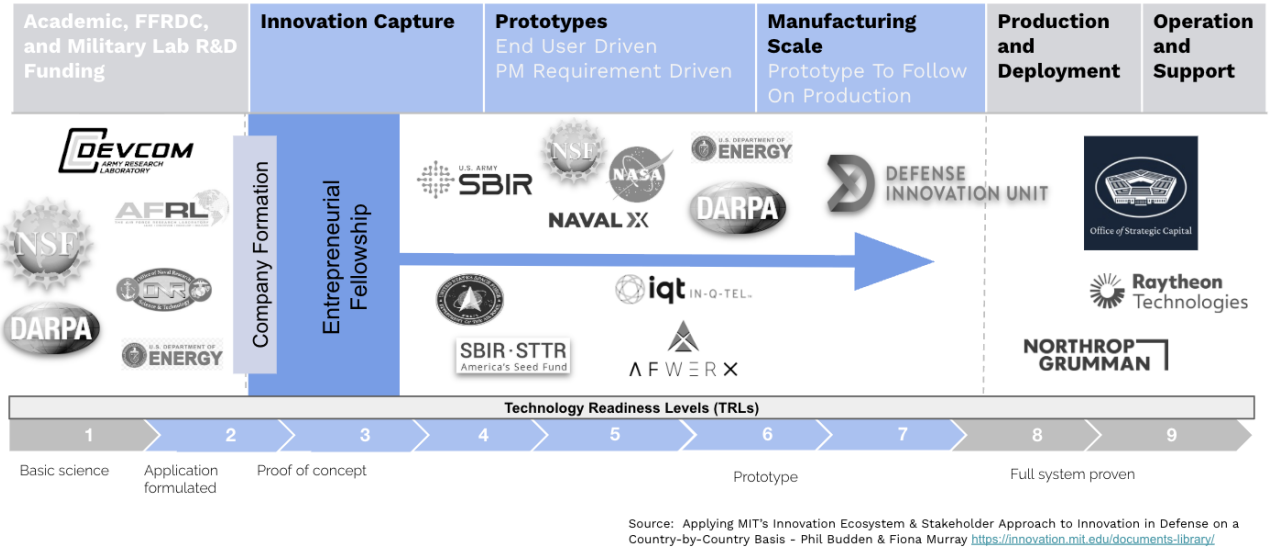
A. Accelerating Transition from Proof-of-Concept to Prototype

A persistent challenge within the U.S. defense innovation system lies not only in identifying promising technologies, but in sustaining them through the highest-risk phase of development. This phase, as previously discussed—often spanning TRLs 2–4—represents one of the highest-risk phases for defense innovation.[55] Technologies that enter defense acquisition channels prematurely—before sufficient technical validation—face increased risk of failure, stalled pilots, or inability to meet performance thresholds under operational conditions.[56] Conversely, premature scaling toward commercial markets can divert technical development away from defense-specific requirements, particularly when venture incentives favor speed over durability.[57] In both cases, the absence of structured support during early-stage development can undermine the likelihood that promising innovations mature into deployable defense capabilities.

Iterative technical validation during the pre-prototype stage is therefore not merely a startup management concern but a strategic consideration for national defense. Advancing from laboratory proof-of-concept to a robust prototype capable of defense testing requires sustained founder engagement, structured technical mentorship, and the time necessary to conduct multiple design-test-refine cycles.[58] Without mechanisms that stabilize founders during this period, the defense innovation pipeline risks either losing promising technologies to attrition or distorting their development trajectory toward non-defense applications.[59]

Entrepreneurial fellowships may represent one institutional response to this structural gap. Participants interviewed in this paper noted that the stipend had been critical to them in enabling full-time work on technology maturation during the highest-risk stage of company formation until SBIR and other defense and venture capital funding became realistically attainable (see Figure 1.)[6061] By providing founder salary support and other resources, such fellowships may reduce early-stage fragility without forcing premature scaling. Rather than accelerating commercialization at all costs, the model emphasizes sustained technical validation and deliberate navigation of acquisition channels.

Figure 1. Defense Invention to Capability Pipeline



Source: Applying MIT's Innovation Ecosystem & Stakeholder Approach to Innovation in Defense on a Country-by-Country Basis - Phil Budden & Fiona Murray https://innovation.mit.edu/documents-library/

B. Reducing Downstream Acquisition Risk

Evidence indicates that entrepreneurial fellowships de-risk downstream DoD acquisition by validating operational relevance early, shaping requirements and concepts of operations before formal programs of record, and improving integration readiness well in advance of the Federal Acquisition Regulation (FAR) pipeline. In contrast to traditional grant mechanisms or late-stage procurement pilots, fellowships intervene upstream, when technologies, teams, and use cases remain malleable. This reduces the likelihood that DoD invests in capabilities that fail to meet operational needs, integrate poorly with legacy systems, or stall during the transition from prototype to production.[626364]

Early Validation of Operational Relevance

A core failure mode in defense innovation is building technically impressive solutions that do not map cleanly onto real operational problems. Performance issues in defense innovation—cost overruns, delivery delays, and fielded systems that become misaligned with evolving missions—are frequently rooted in early-stage misalignment between technical development and operational needs.[65] This misalignment is often compounded when technologies are developed without a clearly defined Concept of Operations (CONOPS). Traditional acquisition processes

often codify requirements years in advance, even as threat environments and technologies continue to evolve. This rigidity increases the likelihood that programs lock into inferior operational and technological choices that are costly to unwind later.[66]

Entrepreneurial fellowships appear to be well positioned to mitigate this risk: by embedding founders with warfighters and defense programs early, providing operational testing and user feedback to accelerate validation of operational relevance well before traditional acquisition timelines, and reducing uncertainty around whether a capability will meaningfully support mission needs.[67] This also allows DoD stakeholders to refine requirements based on demonstrated capabilities rather than abstract forecasts.[68] This upstream co-development reduces the risk of later requirements churn, re-baselining, and costly redesign once programs enter formal acquisition pathways.

Improving Integration Readiness

Integration failures—accreditation hurdles, data interoperability challenges, and platform compatibility issues—often derail transition to production even when prototypes perform well. Furthermore, a RAND study on sources of weapon programs cost growth shows that errors in cost-estimates and late requirements changes, often stemming from unforeseen integration issues, account for 10.1% and 12.9% of cost-growth instances.[69] Entrepreneurial fellowships surface these integration risks early by exposing startups to real-world constraints.[70]

Increasing Capacity for Adhering to Complex Defense Compliance Requirements

In defense-relevant sectors, participants noted that the stipend was particularly helpful in allowing for full-time work on engaging DoD funding offices and complex defense contracting and compliance requirements. By providing salary stipend independent of immediate commercial revenue, entrepreneurial fellowships reduce the need for founders to pursue parallel employment or prematurely seek equity financing. From a defense innovation perspective, this funding directly influences the capacity of startups to engage with defense customers and to navigate multi-stage acquisition processes.[717273]

Cohort-Level Data on DoD Contract Activity Among Fellowship-Supported Ventures

This section provides program data on 41 early-stage ventures participating in an entrepreneurial fellowship that secured at least one contract from across the DoD between 2015 and 2025. Across these ventures, 88 distinct DoD contracts were awarded, representing approximately $48.6 million in total obligated funding. The median contract size was $225,000, consistent with early-stage SBIR, prototype, and pilot-scale engagements. These figures provide a measurable indicator of defense acquisition participation among ventures operating at early technological readiness levels within a defined founder-capitalization environment.[74]

Defense engagement within this cohort was not episodic. The median number of DoD contracts per company was two, indicating that interaction with the acquisition system frequently extended beyond initial entry. Awards were distributed across multiple defense innovation offices and branches including AFWERX, the Air Force Research Laboratory (AFRL), Defense Advanced Research Projects Agency (DARPA) (including multiple program offices), Army, Navy, Space Force, and others demonstrating engagement across research, prototyping, and service-level funding channels.

The temporal distribution of awards is particularly relevant for understanding early-stage defense integration. Of the 88 total contracts, 16 were awarded prior to fellowship participation, 45 were awarded during the fellowship period, and 27 were awarded following completion of the program. Thus, 72 of 88 (91%) contracts were secured during or after structured founder support, totaling $44.04 million. While this analysis does not establish causality, the clustering of awards during and after fellowship participation indicates temporal alignment between early-stage founder capitalization and measurable defense acquisition activity.

The pattern is consistent with the hypothesis that structured founder support may reduce the time required for early-stage ventures to navigate defense funding mechanisms.[75] Although the dataset does not provide a counterfactual comparison, it offers structured descriptive evidence that early-stage founders supported through entrepreneurial fellowships can achieve measurable integration

into defense acquisition channels. This is supported across multiple interviews with founders who reported that participation in an entrepreneurial fellowship enhanced their ability to secure DoD funding by providing financial stability, proposal development support, support on technical validation, and introductions to relevant defense stakeholders.

Case Study: Twelve and U.S. Air Force

An illustrative example of how entrepreneurial fellowships can support the transition of laboratory-stage research toward operational relevance is provided by the synthetic fuel technology company Twelve. Prior to co-founding Twelve, CEO Dr. Etosha Cave participated in the Cyclotron Road entrepreneurial fellowship at Lawrence Berkeley National Laboratory in 2015 funded by the Department of Energy (DOE) through the Lab Embedded Entrepreneurship Program (LEEP).[76]

At the time of participation, Twelve's underlying electrochemical platform—which enables the conversion of carbon dioxide into synthetic fuels—remained at a pre-commercial stage and required sustained experimental refinement.[77] The fellowship environment provided a full salary stipend, funding for technical development, access to laboratory infrastructure, and interdisciplinary mentorship across engineering, commercialization strategy, and systems integration as a bridge between laboratory discovery and applied research, allowing foundational electrochemical processes to be evaluated in the context of real-world deployment constraints.[78]

The operational relevance of such technologies becomes particularly evident when viewed through the lens of military logistics vulnerability. Modern force projection depends heavily on fuel supply chains that must be transported through contested environments. For example, during the height of operations in Afghanistan, attacks on fuel and water convoys were associated with more than 30 percent of U.S. casualties, underscoring the extent to which energy logistics function as a structural operational risk.[79] At the same time, projected fuel demand is expected to increase as advanced sensing systems, autonomous platforms, and distributed operations require greater energy inputs.

Following this period of early technical maturation, the U.S. Air Force partnered with Twelve to assess whether carbon transformation technologies could support forward-deployed fuel production. The collaboration demonstrated the feasibility of synthesizing aviation fuel from captured CO_2 and water using renewable power sources.[80]

The significance of this engagement lies not solely in technological feasibility but in the progression from laboratory-stage research to operationally relevant evaluation. Air Force leadership noted that such systems could enable deployed units to generate fuel in situ, potentially reducing reliance on extended and vulnerable supply chains. Importantly, this trajectory—from early electrochemical research supported through a structured fellowship environment to defense-relevant operational assessment—illustrates how entrepreneurial fellowships may contribute to the maturation of scientific innovations prior to formal acquisition engagement. By enabling sustained technical iteration within an applied context, fellowships can facilitate the translation of emerging research into capabilities that align with operational needs.

Regarding the impact of the fellowship, Dr. Cave noted, “As first-time founders with technical backgrounds, [the program] introduced us to the many dimensions of building and operating an impact-driven company, and in particular it taught us to see things from the eyes of our stakeholders and customers and focus on how to both develop value and communicate value to them effectively.”[81]

Figure 2. Twelve’s aviation fuel plant in Moses Lake, Washington



C. Crowding in Private Capital at the Right Time

Entrepreneurial fellowships also appear to de-risk defense innovation by crowding in private capital at the right stage—after technical and operational risk have been partially retired, but before misaligned incentives distort company trajectories. Timing matters: premature venture capital can push startups toward commercial markets misaligned with defense needs, while late-stage government funding alone often cannot sustain the scale and speed required for competitive advantage.

Early-stage defense-relevant startups face a structural financing gap. Traditional public R&D mechanisms fund projects rather than companies and rarely support team formation, iteration, or commercialization pathways. Meanwhile, private investors are often reluctant to engage early in national security markets due to long sales cycles and regulatory friction.[82] By retiring the highest uncertainty before startups seek venture capital, fellowships improve the probability that private capital reinforces rather than distorts national security outcomes.

Venture capital incentives prioritize speed to scale and large commercial markets, while defense acquisition emphasizes compliance, integration, and long-term sustainment. Without an intermediate de-risking layer, startups are often pushed prematurely toward adjacent commercial markets with faster revenue, diluting defense relevance. From a public-sector perspective, entrepreneurial fellowships improve the efficiency of government capital by enabling small,

early-stage investments across a portfolio of ventures rather than large, late-stage bets on unproven vendors. This portfolio logic mirrors venture investing: many early experiments will fail, but the few that succeed justify the overall investment.

Among the 41 early-stage ventures that received DoD funding during or following participation in an entrepreneurial fellowship, these companies went on to collectively raise approximately $1.27 billion in venture capital—an average of roughly $31 million per company. This data underlines that fellowships may not only be reducing early technical and operational risk but positioning companies to pair government contracts with growth-oriented private capital. By supporting team formation, product maturation, and early defense engagement, fellowships appear to help ventures reach a stage where venture investment can fund organizational build-out, manufacturing readiness, and market expansion—while defense contracts validate use cases and sustain mission relevance. In this way, fellowships could enable a complementary capital stack in which venture funding scales the enterprise and government demand anchor its trajectory in national security needs.

At approximately $350,000 per fellowship, the total amount of direct support across all 41 companies is roughly $14 million. $44.04 million in DoD contract funding and $1.27 billion in private capital secured during or following the fellowship together represent a more than 90x multiple relative to initial direct fellowship support. While this analysis does not establish causality, the magnitude and timing of this relationship are consistent with the hypothesis that modest early-stage investment can help unlock substantially larger downstream public and private capital flows.

It is also worth noting that two of the 41 companies have achieved acquisition outcomes as of the time of publication and, while acquisition is not the primary objective of defense funding, these outcomes provide further evidence of downstream value creation. Given the early-stage, hard-technology focus of these ventures and the various points of company formation between 2015–2025, this level of acquisition activity appears consistent with expected outcomes for comparable early-stage venture portfolios. To the best of the authors' knowledge, the remaining companies from these cohorts continue to be active, though as with any portfolio of early-stage ventures,

some level of attrition is expected given the technical and capital risks inherent at pre-prototype stages.

V. Strategic Technology Domains Where Effects Are Most Pronounced

A. Why Early-Stage Risk Is Domain-Dependent

As briefly discussed earlier in this paper, the structural dynamics described above do not affect all technology domains equally. The consequences of early-stage capital fragility vary substantially depending on technical architecture, manufacturing intensity, and integration complexity. While some defense-relevant sectors and technologies can progress through early development cycles with relatively modest capital and rapid iteration, others face extended validation timelines and capital-intensive maturation pathways. Understanding this domain dependence is essential to evaluating where founder-centric stabilization mechanisms, and consequently entrepreneurial fellowships, may have the greatest impact.

Hardware Versus Software

Software-dominant technologies—particularly those built on modular architectures—often exhibit rapid iteration cycles, low marginal testing costs, and relatively short feedback loops. For example, data analytics tools can frequently reach demonstrable functionality within months. Such ventures are therefore more compatible with traditional venture capital timelines and accelerator models that emphasize rapid product-market fit and scaling.[83]

By contrast, hardware-intensive technologies, including position, navigation, timing (PNT) capabilities, advanced manufacturing, biomanufacturing, advanced materials, microelectronics, and robotics operate under fundamentally different constraints. Prototype development requires fabrication, materials sourcing, laboratory testing, and often specialized equipment. Design iterations are slower and more expensive, and performance validation may depend on environmental conditions that cannot be simulated digitally.[84] As a result, the transition from

proof-of-concept to prototype in hardware domains frequently spans multiple iterative cycles, each requiring incremental capital and sustained founder commitment.

From a defense modernization perspective, many strategically significant domains fall into the latter category.[85] These technologies cannot be meaningfully advanced through software-style acceleration models. Early-stage undercapitalization in hardware domains therefore carries disproportionate strategic risk: promising concepts may fail not because of technical infeasibility, but because the firm lacks sufficient runway to complete validation cycles.

Manufacturing Intensity and Capital Requirements

Closely related to hardware constraints is the issue of manufacturing intensity. Technologies requiring pilot-scale production, precision tooling, or specialized fabrication infrastructure encounter a second layer of early-stage fragility. Even at TRLs 3–5, ventures may need to prove manufacturability, durability, or supply chain viability to qualify for defense integration pathways.[86]

Unlike digital platforms, manufacturing-intensive ventures must navigate supplier relationships, quality assurance protocols, and often regulatory or export-control environments before they are attractive to primes or acquisition offices. These requirements introduce fixed costs that are difficult to amortize at low production volumes. In capital markets optimized for asset-light growth, such ventures are often viewed as high-risk and slow-scaling.[87]

From a national security standpoint, however, manufacturing capability is not incidental—it is central to force readiness and industrial base resilience. The erosion of domestic manufacturing capacity has been repeatedly identified as a strategic vulnerability.[88] Early-stage technologies intended to restore or reinforce domestic manufacturing capabilities therefore face a paradox: they are strategically valuable but capital-intensive at inception. Founder stabilization mechanisms may be particularly consequential in such domains because they allow iterative validation of manufacturing processes without immediate pressure to achieve commercial-scale throughput.

Domain Implications for Early-Stage Stabilization

Taken together, these domain-dependent dynamics suggest that the effects of early-stage stabilization mechanisms are unlikely to be uniform across the defense innovation landscape. In software-centric sectors with rapid iteration cycles, capital fragility may be less acute, and traditional venture-backed acceleration models may suffice. In hardware-intensive, manufacturing-dependent, and integration-complex domains, however, the mismatch between capital timelines and technical maturation is more pronounced.

If modernization priorities increasingly emphasize resilient supply chains, advanced materials, energy storage, autonomy in contested environments, and next-generation manufacturing capabilities, then early-stage stabilization becomes strategically consequential. These domains require sustained technical iteration before scaling, and the failure mode is often premature attrition rather than technical impossibility.

Entrepreneurial fellowships, by design, may intervene at precisely this high-risk intersection. By extending founder runway, supporting iterative validation, and deferring premature scaling pressures, such mechanisms may exert disproportionate influence in domains where early-stage capital misalignment is most acute. The empirical patterns presented earlier—particularly in hardware- and manufacturing-intensive ventures engaging multiple defense components—are consistent with this domain-sensitive hypothesis.

This domain-based framing reframes the question from whether entrepreneurial fellowships are broadly beneficial to where they are most strategically relevant. If early-stage risk is unevenly distributed across technology sectors, then policy responses should similarly be targeted. Stabilization mechanisms may be most justified not in areas already well-served by venture capital, but in strategically critical domains where capital market mismatch and integration complexity converge.

B. Priority Domains for Non-Dilutive Grants and Entrepreneurial Fellowships

The preceding analysis established that early-stage stabilization mechanisms may yield the greatest impact in sectors characterized by hardware intensity, manufacturing complexity, long iteration cycles, and integration risk. This section identifies specific domains where non-dilutive grants and entrepreneurial fellowships are most strategically consequential for defense and national security.

The DoD recently revised its Critical Technology Areas (CTAs), announced in November 2025 by Under Secretary of Defense for Research and Engineering Emil Michael. By consolidating the previous fourteen CTAs into six focused priorities including Applied Artificial Intelligence (AAI), Biomanufacturing (BIO), Contested Logistics Technologies (LOG), Quantum and Battlefield Information Dominance (Q-BID), and Scaled Hypersonics (SHY)—the Department signaled an intent to deliver capabilities within twelve-to-thirty-six-month "sprints."[89]

Yet the urgency of these sprints presupposes a pipeline of technically mature ventures across all priority domains ready to transition into rapid development—precisely the pipeline that entrepreneurial fellowships appear to be designed to create. Without sustained early-stage investment in these foundational domains, sprint timelines risk being constrained not by program management velocity but by the absence of viable technology entrants. Entrepreneurial fellowships and non-dilutive grants could help stabilize ventures in these critical technologies during the highest-risk development phase and increase the supply of technologies for the Department's accelerated CTA delivery model.

Position, Navigation, and Timing (PNT)

The U.S. military's dependence on GPS for positioning, navigation, and timing represents a widely recognized yet inadequately addressed vulnerability. GPS signals underpin precision-guided munitions, command-and-control synchronization, and nearly every networked military system, yet they are susceptible to jamming, spoofing, and physical attack.[90] Alternative PNT technologies—quantum inertial navigation sensors, low Earth orbit (LEO) satellite-based timing, enhanced long-range navigation (eLoran), and signals-of-opportunity receivers—require

specialized hardware development in atomic physics and precision optics, validation in controlled electromagnetic environments, and complex system-of-systems integration across service branches.[91]

While the Defense Innovation Unit and SpaceWERX have begun soliciting proposals for alternative PNT prototypes, these efforts remain downstream of the earliest company formation stages.[92] Fellowships would intervene where academic research in quantum sensing, and alternative navigation is ready for translation into company-led prototype development but lacks the organizational and financial infrastructure to survive iterative hardware validation. PNT intersects directly with the Q-BID critical technology area and underpins operational effectiveness across the CTAs.

Advanced Manufacturing (Including Biomanufacturing)

The Department's designation of Biomanufacturing (BIO) as a Critical Technology Area underscores the strategic urgency of advanced manufacturing. Biomanufacturing offers the potential to produce critical chemicals, fuels, and materials through engineered biological systems rather than petrochemical processes, reducing dependence on vulnerable foreign supply chains.[93] The National Security Commission on Emerging Biotechnology (NSCEB) has warned that without adequate early-stage support, the Department risks losing valuable industry partnerships before they reach production readiness.[94] Fellowships in this domain would help bridge the gap in biomanufacturing and contested logistics critical technology areas.

Advanced Materials

Advanced materials—novel composites, high-temperature alloys, metamaterials, energetic materials, and next-generation structural systems—constitute a foundational enabling technology across the CTA portfolio. Hypersonic vehicles require thermal protection materials for Mach 5+ flight; directed energy systems demand novel optical and thermal management materials; and next-generation platforms depend on lightweight composites and advanced coatings. Breakthroughs in this domain have multiplicative effects, directly enabling progress in SHY,

SCADE, and LOG. At the same time, materials ventures face among the longest development cycles of any technology sector, requiring years of iterative synthesis, characterization, environmental testing, and manufacturing process development.

VII. Policy Recommendations: Institutionalizing Early-Stage Fellowship Funding

The preceding analysis identifies a structural gap in the U.S. defense innovation ecosystem: the absence of sustained, institutionalized support for researchers and entrepreneurs at the earliest pre-commercial stages, before technologies are mature enough for SBIR grants, before companies are formed enough for venture investment or DIU prototype agreements, and before concepts are suitable for traditional acquisition pathways. Addressing this gap requires policy interventions that are realistic, implementable within existing authorities, and complementary to existing defense innovation organizations. The following recommendations outline how early-stage fellowship funding can be institutionalized as a durable instrument of U.S. defense innovation strategy, with success measured primarily by transitions into existing DoD mechanisms such as SBIR, DIU, and service programs.[95]

A. Expand Early-Stage Fellowship Funding Within DoD

The DoD may consider assessing the feasibility of establishing a dedicated early-stage fellowship program to provide non-dilutive grants to scientists and engineers working on defense-relevant technologies. This program could function as an upstream complement to existing mechanisms, filling the gap between basic research and the earliest commercial-stage interventions. Unlike SBIR, which funds project proposals tied to predefined topics and commercialization plans, fellowships could fund people and teams during the pre-proposal and pre-requirements phase, when use cases, technical direction, and founding teams are still being formed.[96]

With respect to funding, fellowship awards of approximately $350,000 total for at least two years would provide founder salary support, research and development funding, and structured milestone guidance during the transition from proof of concept to company formation at TRLs 2-

6, where this paper has established defense innovation risk is highest. Funding should be non-dilutive to avoid premature incentive distortions and allow founders to pursue defense-relevant pathways without pressure to pivot toward faster commercial markets. Selection should prioritize DoD CTAs, specifically hardware-intensive domains where private capital is most misaligned.[97] A pilot cohort of 20-30 fellows per year over two years would represent a negligible fraction of the DoD RDT&E budget ($7 million -$10.5 million in direct fellowship funding annually) but could materially expand the upstream pipeline feeding SBIR, DIU, and service acquisition programs.

The program could be coordinated by the Office of the Under Secretary of Defense for Research and Engineering (USD (R&E)), with fellows aligned to DARPA or the service research labs' science and technology priorities. Existing authorities, including OTA or Broad Agency Announcement (BAA) are sufficient to launch such a program without new legislation.[98] Over time, legislative authorization may help institutionalize the fellowship and provide long-term funding stability. To avoid perpetual pilot status, the fellowship should be authorized initially as a five-year budget line with pre-committed scale milestones, contingent on transition metrics into downstream programs.

B. Align Fellowships with Defense Stakeholders Without Gatekeeping

At a minimum to increase their chance of success, fellowships should integrate early exposure to defense problem sets while avoiding premature bureaucratic gatekeeping. Engagement with requirements owners and operators should occur early and iteratively, shaping technical development before designs harden. However, fellows should not be subjected to formal acquisition requirements or program office approval processes at the proof-of-concept stage, which would undermine exploratory development. Engagement with program offices should be brokered by fellowship administrators in order to avoid imposing additional workload on already constrained program offices. The fellowship should also introduce this interaction earlier, at the proof-of-concept phase, when iteration is least costly and learning is fastest, and could include workshops and speaker events with DoD personnel, defense primes and dual use defense investors, as well as provide introductions to defense funding/ contracting.

The fellowship should provide structured bridges to SBIR solicitations, DIU prototype opportunities, and defense-oriented investors, alongside support for CONOPS development and transition planning. Program performance should be evaluated on portfolio-level outcomes rather than individual project success, recognizing that early-stage attrition is a feature of risk absorption rather than a failure of program design. Without institutionalized handoffs into acquisition and investment pipelines, fellowship programs risk producing technically capable ventures that nonetheless fail to translate into sustained defense adoption.[99]

C. Improve Interagency Coordination

Lastly, this paper recommends improving coordination between the DoD and civilian science agencies to better leverage federally funded basic research investments. Agencies such as DOE and NSF support early-stage companies through programs such as SBIR and entrepreneurial fellowships with defense relevance. Fellows supported by DOE LEEP and NSF TIP entrepreneurial fellowship programs have gone on to secure DoD funding after using fellowship support to stabilize their company and mature their technologies to a level suitable for defense programs with Air Force, Navy, Army, and more, as previously noted.[100][101][102]

Improved interagency alignment could allow DoD to benefit more systematically from civilian R&D investments without assuming responsibility for absorbing early-stage technical and organizational risk alone. Coordinated support for entrepreneurial fellowship models—building on prior NSF efforts—would help extend technical runways, facilitate early customer discovery, and strengthen organizational readiness among research-originating ventures. By linking civilian research investments with defense-oriented transition pathways, the federal government can improve continuity across the innovation pipeline and reduce the likelihood that strategically significant technologies stall between discovery and acquisition relevance.

VIII. Conclusion: Early-Stage Funding as Strategic Infrastructure

This paper has argued that early-stage entrepreneurial fellowships can function as strategic infrastructure within the defense innovation pipeline. By stabilizing founders before venture readiness or formal acquisition engagement, these programs address a structural gap that existing mechanisms such as SBIR, venture capital, and other accelerator programs are not designed to fill. This founder-capital stabilization supports sustained technical development, early engagement with defense stakeholders, and the organizational maturation required to navigate acquisition pathways.

The empirical patterns examined here suggest that ventures supported through structured fellowship models can achieve measurable integration into defense funding channels across multiple stages of engagement. While this analysis does not establish causality, the recurrence and timing of contract activity are consistent with the hypothesis that early stability reduces barriers that often prevent promising technologies from progressing beyond initial proof-of-concept funding.

As technological competition intensifies and modernization timelines shorten, the strategic question is no longer whether the U.S. can invent first, but whether it can field first, and early-stage funding should be viewed as the enabling infrastructure to do so. Entrepreneurial fellowship programs that provide multi-year stability and structure may help ensure that scientifically promising technologies are not lost prior to reaching acquisition relevance, strengthening the continuity of the defense industrial base.

Endnotes

[1] Lisa A. Aronsson, *Transatlantic Perspectives on Defense Innovation: Issues for Congress* (Washington, DC: Congressional Research Service, 2018), 1–5.
[2] U.S. Congress, *Consolidated Appropriations Act, 2026*, H.R. 7148, 119th Cong., 2nd Sess. (2026), Division A, Title IV (Research, Development, Test, and Evaluation), p. 19.
[3] Carlton Haelig and Philip Sheers, *Stuck in the Cul-de-Sac: How U.S. Defense Spending Prioritizes Innovation over Deterrence* (Washington, DC: Center for a New American Security, October 21, 2025), 11. https://www.cnas.org/publications/reports/stuck-in-the-cul-de-sac.
[4] John F. Sargent Jr. and Marcy E. Gallo, *The Global Research and Development Landscape and Implications for the Department of Defense*, R45403 (Washington, DC: Congressional Research Service, 2021), https://www.congress.gov/crs-product/R45403.
[5] U.S. Government Accountability Office, *Weapon Systems Annual Assessment: Report to Congressional Committees*, GAO-25-107569 (Washington, DC: U.S. Government Accountability Office, June 11, 2025).
[6] U.S. Government Accountability Office, *Defense Technology Development: Transition of DOD-Funded Technologies from Science and Technology Programs*, GAO-14-748T (Washington, DC: U.S. Government Accountability Office, September 2014).
[7] Aronsson, *Transatlantic Perspectives*, 21–25.
[8] U.S. Department of Defense, *2022 National Defense Strategy of the United States of America* (Washington, DC: DoD, 2022).
[9] Aronsson, *Transatlantic Perspectives*, 21–25.
[10] Government Accountability Office, *Weapon Systems Annual Assessment: Report to Congressional Committees*, GAO-25-107569.
[11] Haelig and Sheers, 12.
[12] Defense Innovation Board, *Terraforming the Valley of Death: Making the Defense Market Navigable for Startups* (Washington, DC: U.S. Department of Defense, July 17, 2023), https://innovation.defense.gov/Portals/63/DIB_Terraforming%20the%20Valley%20of%20Death_230717_1.pdf.
[13] Aronsson, *Transatlantic Perspectives*, 1–5.
[14] U.S. Department of War, *2026 National Defense Strategy* (Washington, DC, 2026).
[15] U.S. Small Business Administration, *SBIR/STTR Program Policy Directive* (Washington, DC: U.S. Small Business Administration, 2023).
[16] U.S. Department of Defense, Office of Small Business Programs, *DoD SBIR/STTR Program Overview* (Washington, DC: Department of Defense, 2022).
[17] U.S. Government Accountability Office, *Defense Innovation: DOD Needs to Improve How It Communicates and Measures the Performance of Its Innovation Efforts*, GAO-23-106089 (Washington, DC: U.S. Government Accountability Office, 2023).
[18] U.S. Small Business Administration, *SBIR/STTR Program Policy Directive* (Washington, DC: U.S. Small Business Administration, 2023); U.S. Government Accountability Office, *Small Business Research Programs: Actions Needed to Improve Commercialization Outcomes*, GAO-17-329 (Washington, DC: U.S. Government Accountability Office, 2017).
[19] National Academies of Sciences, Engineering, and Medicine, *An Assessment of the SBIR and STTR Programs at the Department of Defense* (Washington, DC: National Academies Press, 2019).
[20] William R. Kerr and Ramana Nanda, "Financing Innovation," *Annual Review of Financial Economics* 7 (2015): 445–462.
[21] National Academies of Sciences, Engineering, and Medicine, *An Assessment of the SBIR and STTR Programs at the Department of Defense* (Washington, DC: National Academies Press, 2019).
[22] U.S. Government Accountability Office, *Defense Acquisitions: Assessments of Major Weapon Programs*, annual reports (Washington, DC: U.S. Government Accountability Office, various years); U.S. Government Accountability Office, *Small Business Research Programs: Actions Needed to Improve Commercialization Outcomes*, GAO-17-329 (Washington, DC: U.S. Government Accountability Office, 2017).
[23] U.S. Department of Defense, *2022 National Defense Strategy of the United States of America* (Washington, DC: Department of Defense, 2022), https://media.defense.gov/2022/Oct/27/2003103845/-1/-1/1/2022-NATIONALDEFENSE-STRATEGY-NPR-MDR.PDF.

[24] Paul A. Gompers and David Lane, *SRS and the Defense Innovation Unit: Rethinking Procurement* (9-220-047) (Boston: Harvard Business School, 2020).
[25] U.S. Government Accountability Office, *Defense Innovation: DOD Needs to Improve How It Communicates and Measures the Performance of Its Innovation Efforts*, GAO-23-106089 (Washington, DC: U.S. Government Accountability Office, 2023).
[26] Carlos Martí Sempere, "A Survey of Performance Issues in Defence Innovation," *Defence and Peace Economics* 28, no. 3 (2017): 319–343.
[27] National Security Commission on Artificial Intelligence, *Final Report* (Washington, DC: National Security Commission on Artificial Intelligence, 2021), sections on private capital in defense innovation; U.S. Department of Defense, Defense Innovation Unit, *Annual Report* (Washington, DC: Department of Defense, various years).
[28] Ajay Agrawal, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence* (Boston: Harvard Business Review Press, 2018), on capital intensity and scaling dynamics; National Academies of Sciences, Engineering, and Medicine, *Securing Advanced Manufacturing in the United States* (Washington, DC: National Academies Press, 2021).
[29] U.S. Government Accountability Office, *Defense Acquisitions: Assessments of Major Weapon Programs*, annual reports (Washington, DC: U.S. Government Accountability Office, various years).
[30] Paul Gompers and Josh Lerner, *The Venture Capital Cycle*, 2nd ed. (Cambridge, MA: MIT Press, 2004).
[31] Dan Breznitz, *Innovation in Real Places: Strategies for Prosperity in an Unforgiving World* (Oxford: Oxford University Press, 2021); National Security Commission on Artificial Intelligence, *Final Report* (Washington, DC: National Security Commission on Artificial Intelligence, 2021).
[32] Tai Ming Cheung, *Innovate to Dominate: The Rise of the Chinese Techno-Security State* (Ithaca, NY: Cornell University Press, 2022); Matthew Weinzierl and Brendan Rosseau, *The United States National Security Apparatus, Multipolarity, and the Rise of Commercial Space* (Boston: Harvard Business School, 2022).
[33] Dana J. Pernin et al., *Maintaining the Competitive Advantage: A Strategy for Defense Innovation* (Santa Monica, CA: RAND Corporation, 2019).
[34] Gompers and Lane, 4-5.
[35] Gompers and Lane, 4-5.
[36] *The Economist*, "How China Became an Innovation Powerhouse," August 25, 2025, https://www.economist.com/business/2025/08/25/how-china-became-an-innovation-powerhouse
[37] U.S. Department of Defense, *Technology Readiness Assessment (TRA) Guidance* (Washington, DC: Department of Defense, 2011).
[38] Paul Gompers and Josh Lerner, *The Venture Capital Cycle*, 2nd ed. (Cambridge, MA: MIT Press, 2004).
[39] U.S. National Science Foundation, "NSF Launches Entrepreneurial Fellowship for Engineers and Scientists," September 19, 2022, NSF Launches Entrepreneurial Fellowship for Engineers and Scientists.
[40] U.S. National Science Foundation, "NSF Launches Entrepreneurial Fellowship for Engineers and Scientists" (Alexandria, VA: National Science Foundation, September 19, 2022), https://www.nsf.gov/tip/updates/nsf-launches-entrepreneurial-fellowship-engineers
[41] Ibid.
[42] Ibid.
[43] Ibid.
[44] Eric Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (New York: Crown Business, 2011); William R. Kerr and Ramana Nanda, "Financing Innovation," *Annual Review of Financial Economics* 7 (2015): 445–462.
[45] David Mackanic, CEO and Founder, Anthro Energy, email correspondence with the author, February 2026.
[46] Jae Cho, Founder and CEO, Enertia Microsystems, email correspondence with the author, February 2026.
[47] Blake Herren, Founder and CEO, Raven Space Systems, email correspondence with the author, February 2026.
[48] Ibid.
[49] U.S. National Science Foundation, "NSF Launches Entrepreneurial Fellowship for Engineers and Scientists" (Alexandria, VA: National Science Foundation, September 19, 2022), https://www.nsf.gov/tip/updates/nsf-launches-entrepreneurial-fellowship-engineers.
[50] U.S. Small Business Administration, *SBIR and STTR Program Overview* (Washington, DC: Small Business Administration, n.d.), https://www.sbir.gov/about.
[51] National Academies of Sciences, Engineering, and Medicine, *An Assessment of the SBIR Program at the National Institutes of Health* (Washington, DC: National Academies Press, 2015).
[52] Susan G. Cohen, Daniel Fehder, Yael V. Hochberg, and Fiona Murray, "The Design of Startup Accelerators," *Research Policy* 48, no. 7 (2019): 1781–1797.

[53] U.S. Economic Development Administration, *Incubating Success: Incubation Best Practices That Lead to Successful New Ventures* (Washington, DC: U.S. Department of Commerce, 2011).
[54] U.S. National Science Foundation, "NSF Launches Entrepreneurial Fellowship for Engineers and Scientists" (Alexandria, VA: National Science Foundation, September 19, 2022), https://www.nsf.gov/tip/updates/nsf-launches-entrepreneurial-fellowship-engineers
[55] U.S. Department of Defense, Office of the Under Secretary of Defense for Research and Engineering, *Defense Innovation Ecosystem Report* (Washington, DC: Department of Defense, 2019).
[56] U.S. Government Accountability Office, *Defense Acquisitions: Assessments of Major Weapon Programs*, annual reports (Washington, DC: U.S. Government Accountability Office, various years).
[57] Dan Breznitz and Michael Murphree, *Run of the Red Queen: Government, Innovation, Globalization, and Economic Growth in China* (New Haven: Yale University Press, 2011).
[58] Eric Ries, *The Lean Startup: How Today's Entrepreneurs Use Continuous Innovation to Create Radically Successful Businesses* (New York: Crown Business, 2011).
[59] National Security Commission on Artificial Intelligence, *Final Report* (Washington, DC: National Security Commission on Artificial Intelligence, 2021)
[60] Blake Herren, Founder and CEO, Raven Space Systems, email correspondence with the author, February 2026.
[61] David Mackanic, CEO and Founder, Anthro Energy, email correspondence with the author, March 2026.
[62] Sempere, 319–343. Gompers and David Lane, 3–5.
[63] Gompers and Lane, 3–5.
[64] Joshua Lev Krieger and Josh Lerner, *In-Q-Tel: Innovation on a Mission* (Boston: Harvard Business School, 2022), 4–9.
[65] Sempere, 319-343.
[66] Sempere, 319-343.
[67] Gompers and Lane, 3-5.
[68] Gompers and Lane, 4.
[69] J. G. Bolten, R. S. Leonard, M. V. Arena, A. Younossi, and J. M. Sollinger, *Sources of Weapon System Cost Growth: Analysis of 35 Major Defense Acquisition Programs* (Santa Monica, CA: RAND Corporation, 2008).
[70] Gompers and Lane, 5.
[71] Dan Oran, Founder and CEO, Irradiant Technologies, email correspondence with the author, February 2026.
[72] Blake Herren, Founder and CEO, Raven Space Systems, email correspondence with the author, February 2026.
[73] David Mackanic, CEO and Founder, Anthro Energy, email correspondence with the author, March 2026.
[74] Program data provided by Activate, covering fellowship participants between 2015 and 2025.
[75] Program data provided by Activate, covering fellowship participants between 2015 and 2025.
[76] Lawrence Berkeley National Laboratory, *10 Years of Cyclotron Road Helping Entrepreneurs Bring Technologies to Market*, January 14, 2026, https://newscenter.lbl.gov/2026/01/14/10-years-of-cyclotron-road-helping-entrepreneurs-bring-technologies-to-market/
[77] Etosha Cave, CEO of Twelve, interview
[78] Lawrence Berkeley National Laboratory, *10 Years of Cyclotron Road Helping Entrepreneurs Bring Technologies to Market*, January 14, 2026, https://newscenter.lbl.gov/2026/01/14/10-years-of-cyclotron-road-helping-entrepreneurs-bring-technologies-to-market/
[79] Corrie Poland, "The Air Force Partners with Twelve, Proves It's Possible to Make Jet Fuel out of Thin Air," U.S. Air Force, October 22, 2021, https://www.af.mil/News/Article-Display/Article/2819999/.
[80] Ibid.
[81] Etosha Cave, CEO of Twelve, interview
[82] Joshua Lev Krieger and Josh Lerner, *In-Q-Tel: Innovation on a Mission* (9-823-031) Boston: Harvard Business School, 2022.
[83] Nicholas Bloom, Charles I. Jones, John Van Reenen, and Michael Webb, "Are Ideas Getting Harder to Find?" *American Economic Review* 110, no. 4 (2020): 1104–1144.
[84] Willy Shih, "What It Takes to Reshore Manufacturing Successfully," *MIT Sloan Management Review* 56, no. 2 (2014).
[85] U.S. Department of Defense, *Critical and Emerging Technologies List Update* (Washington, DC: DoD, 2023).
[86] U.S. Government Accountability Office, *Technology Readiness Assessment Guide: Best Practices for Evaluating the Readiness of Technology for Use in Acquisition Programs and Projects* (GAO-20-48G, Washington, DC: GAO, 2020).

[87] David Adler "Financing Advanced Manufacturing: Why VCs Aren't the Answer," *American Affairs* 3, no. 2 (Summer 2019), https://americanaffairsjournal.org/2019/05/financing-advanced-manufacturing-why-vcs-arent-the-answer/.
[88] Becca Wasser and Philip Sheers, *From Production Lines to Front Lines: Revitalizing the U.S. Defense Industrial Base for Future Great Power Conflict* (Washington, DC: Center for a New American Security, April 2025), https://www.cnas.org/publications/reports/from-production-lines-to-front-lines.
[89] Under Secretary of War for Research and Engineering Emil Michael, "Six Critical Technology Areas for the War Department" (Washington, DC: Department of Defense, November 17, 2025); Lindsay McKenzie, "Department of Defense Narrows R&D Priorities List," AIP FYI, November 19, 2025; OUSW(R&E), "Senior Officials Announced for the War Department's Six Critical Technology Areas," January 29, 2026.
[90] Government Accountability Office, *GPS Disruptions Could Have Significant Effects on Military Operations* (GAO-22-104629, 2022)
[91] National Academies of Sciences, *A Review of the Department of Defense's Approach to Resilient Positioning, Navigation, and Timing* (2021)
[92] Defense Innovation Unit and SpaceWERX have issued solicitations for alternative and resilient PNT prototype development through R-PNT and Assured PNT innovation programs.
[93] Department of Defense, *Biomanufacturing Strategy* (2023), highlighting engineered biology as a pathway to secure domestic production of fuels, materials, and chemicals currently reliant on foreign supply chains.
[94] National Security Commission on Emerging Biotechnology, *Interim Report* (2024)
[95] Government Accountability Office, *Defense Innovation: DOD Needs to Improve How It Communicates and Measures the Performance of Its Innovation Efforts*, GAO-23-106089 (Washington, DC: GAO, 2023).
[96] U.S. Small Business Administration, *SBIR/STTR Program Policy Directive* (Washington, DC: SBA, 2023).
[97] U.S. Department of War, "Critical Technology Areas for the Department of Defense," Office of the Under Secretary of Defense for Research and Engineering (Washington, DC: DoW, 2025).
[98] 10 U.S.C. § 4022; Gompers and Lane, 4–5.
[99] Government Accountability Office, GAO-20-439, 2020.
[100] Blake Herren, Founder and CEO, Raven Space Systems, email correspondence with the author, February 2026.
[101] Corrie Poland, "The Air Force Partners with Twelve, Proves It's Possible to Make Jet Fuel out of Thin Air," U.S. Air Force, October 22, 2021, https://www.af.mil/News/Article-Display/Article/2819999/.
[102] Lawrence Berkeley National Laboratory, *10 Years of Cyclotron Road Helping Entrepreneurs Bring Technologies to Market*, January 14, 2026, https://newscenter.lbl.gov/2026/01/14/10-years-of-cyclotron-road-helping-entrepreneurs-bring-technologies-to-market/

AI Without Authority: Workforce Governance as the Limiting Factor in National Security Innovation

Desiree Lorell

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Authors Bio

Desiree Lorell is a doctoral researcher in Instructional Technology at the American College of Education. Her work focuses on information and communication technology governance in national security organizations. Her doctoral research examines how DoD stakeholders select and implement ICT in learning and training environments, emphasizing on governance, adoption constraints, and mission outcomes. She is an inaugural AI & National Security Convergence Fellow at the Council on Strategic Risks and the Future of Life Institute. Her work seeks to bridge academic analysis and practitioner insight to inform responsible technology adoption in complex security systems.

Abstract

Artificial intelligence (AI) is rapidly being integrated into national security organizations, promising speed, efficiency, and enhanced decision-making. Yet AI adoption often prioritizes tools and technical capability over workforce and governance systems responsible for interpreting and validating and acting on AI-enabled outputs. Drawing on patterns identified through doctoral research on information and communication technology selection implementation in defense learning and operational environments, and insights from a national security AI fellowship. Examines how AI adoption reshapes epistemic authority inside organizations. Unlike prior "game-changing" technologies, AI produces answer-like outputs compressing deliberation timelines and increasing pressure on judgment. Without explicit training, doctrine, and workforce safeguards, this dynamic can erode accountability, distort readiness, and incentivize premature adoption. The analysis identifies three failure modes. First, workforce systems often lack the data fidelity and role clarity to determine who is qualified to rely on or contest AI outputs. Second, governance mechanisms are often introduced after acquisition, limiting their influence on system design and workforce preparation. Third, productivity narratives emphasize efficiency gains while underrepresenting the coordination and verification burden that emerge in operational contexts. This study advances a workforce-centered governance model for AI integration in national security. It emphasizes tiered AI exposure based on readiness, doctrinal separation between automation, decision support, and authority, and institutionalized "check-the-checker" norms that preserve human judgment and command accountability. By reframing AI adoption as a workforce and governance challenge, this study offers a practical path for aligning innovation with mission assurance in contested and high-stakes environments.

Keywords: AI governance; mission assurance; workforce readiness; automation bias; decision authority; verification; institutional trust

Introduction

In 2025, national-level policy framing positioned artificial intelligence (AI) as a strategic capability essential to United States (U.S.) competitiveness and national security, emphasizing accelerating innovation and enabling productivity-enhancing uses at scale.[1] The White House's *America's AI Action Plan* explicitly states that "the U.S. is in a race to achieve global dominance in artificial intelligence (AI)," thus, framing AI leadership as consequential for economic competitiveness and national security outcomes.[1] According to the action plan, AI capacity is linked to national advantage. It asserts that "those possessing the largest AI ecosystem will reap broad economic and military benefits." This reflects leaders' expectations that AI will deliver a strategic advantage through faster and more capable decision-making, effective execution, and rapid innovation across sectors.[1]

This sense of urgency is reinforced through White House communication accompanying the plan, clearly stating, "To remain the leading economic and military power, the U.S. must win the AI race."[2] In the plan, federal action centers on few pillars, including accelerating innovation, building national infrastructure to support AI at scale, and leading international diplomacy and security.[1]

At the operational level, this policy direction is reinforced through recent AI acceleration strategy. A department-level strategy memo characterizes the transformation explicitly as "a race," emphasizing that the pace of commercial AI innovation requires rapid institutional adaptation.[3] The speed imperative is unambiguous as the memo categorically states that "speed wins" and directs leaders to treat "cycle time and adoption rates as decisive variables" in AI capability development and scaling.[3] Consistent with traditional defense advantage narratives,

the memo contends that AI must be used to make warfighters “more lethal and efficient”, directing the department toward emerging as an “AI-first warfighting force.” [3]

How does the acceleration of AI shape national security decision-making processes in the United States? Organizations are being pushed to adopt AI rapidly as governance systems processes and workforce readiness continue to develop in parallel.[13] The strategy memo calls for eliminating “bureaucratic barriers” that delay adoption, including friction with authorizations, testing, and compliance processes, while simultaneously acknowledging the need for foundational enablers such as policies and talent.[3]

These pressures collectively reshape how knowledge judgment and authority circulate inside national security organizations. As AI-enabled systems are increasingly positioned as tools for speed efficiency and advantage[13] their outputs are rapidly entering decision processes as inputs. In the absence of clearly defined workforce roles training pathways and governance mechanisms[789] this dynamic creates conditions that can lead to institutional overtrust and the blurring of accountability boundaries.

Workforce governance that converts AI capability into authorized, accountable action is the limiting factor in national security AI innovation. AI can swiftly generate recommendations; however, speed is advantageous only when outputs can be verified, contested, and adjudicated by qualified personnel operating within clear authority boundaries. When role qualification, escalation rules, and verification norms are absent or inconsistent, organizations either slow its adoption to manage risk informally or accelerate adoption and absorb hidden risk through institutional overtrust. In both cases, governance capacity is the determining factor in whether AI can scale as a mission asset.

The primary national security risk associated with near-term AI is institutional overtrust driven by insufficient workforce readiness, unclear authority boundaries, and misaligned governance, rather than autonomous behavior or machine agency. Practically, near-term AI will not undermine national security by acting on its own, but by reshaping decision systems in subtle, incremental ways. The risk being that organizations, both knowingly and unknowingly, transfer epistemic authority to AI outputs as personnel get overwhelmed, policies lag behind implementation, and accountability becomes diffuse.

Despite growing research on AI in national security, studies have primarily focused on system capability performance and responsible AI principles at a policy level. Human factors research has examined automation bias and overreliance while governance frameworks emphasize lifecycle risk management and ethical use. However, these approaches do not fully account for how authority accountability and workforce structure are reshaped when AI outputs enter decision processes and limited attention has been given to how workforce readiness role clarity and adjudication mechanisms determine whether AI-enabled systems can be used reliably in high-consequence environments.

This study addresses the lack of attention to how workforce readiness role clarity and adjudication mechanisms shape authority and accountability in AI-enabled decision processes by advancing a workforce-centered governance approach. This study is divided into five sections. First it explains how contemporary AI systems differ from earlier decision-support technologies by producing outputs in the form of answers thereby reshaping how epistemic authority forms within organizations and second it employs human factors research on automation bias and overreliance to explain why answer-like output increases the likelihood of institutional overtrust particularly when facing time pressures and third it analyzes three failure modes that emerge

when AI adoption outpaces workforce readiness and governance and fourth it advances a workforce-centered governance model built on tiered AI exposure explicit separation between automation decision support and authority and institutionalized verification and escalation norms designed to preserve accountability in AI-enabled decision processes and finally it offers a first-year implementation roadmap to help inform leaders on how to align AI adoption with mission assurance in high-stakes environments.

What Makes AI Different From Past Tools

AI differs from prior decision-support technologies in how its outputs are perceived and used. Traditional tools, such as databases, dashboards, and analytic models function as instruments that require interpretation, synthesis, and explicit human judgment. By contrast, many contemporary AI systems generate outputs in the form of answers. They summarize, recommend, predict, or rank options in ways that appear coherent, complete, and authoritative, often without exposing underlying uncertainty, assumptions, or trade-offs.

This answer-like output has important organizational consequences. When outputs are presented as finished products, they compress deliberation timelines, thus reducing opportunities for questioning, debating, and adjudication. In time-constrained environments, AI outputs can function as cognitive shortcuts, substituting for deliberate forms of collective reasoning. These systems can narrow decision pathways by implicitly framing what counts as relevant information and which options merit consideration.

Unlike earlier “game-changing” technologies, including databases dashboards and statistical models, AI acts as a decision-making tool, as well as a team member for forming judgments and adjudicating decisions. Its outputs are often considered as epistemic claims as opposed to the provisional aids that they truly are. This distinction is significant because

organizational authority is closely associated with production and validation of knowledge and with who is empowered to challenge it. When AI systems generate outputs in the form of answers, they can subtly acquire epistemic weight without corresponding changes to doctrine, training, or accountability structures.

"Answer-Like" Output and Overtrust

The tendency to overtrust automated systems is well documented in human factors research. Studies on automation bias indicate that individuals frequently defer to automated recommendations even when these conflict with their own judgment or with available evidence.[4,5] This effect is not limited to novice users; even experienced operators and highly trained teams are susceptible, particularly when operating under time constraints or when they perceive systems as highly reliable.

Overtrust manifests in identifiable behavioral patterns. Users may neglect seeking corroborating information or discount contradictory cues therefore accepting AI-generated outputs with minimal verification. In team settings AI recommendations can anchor discussions and shape subsequent interpretations narrowing the range of perspectives considered. Once an automated output is introduced alternative hypotheses are often evaluated in relation to it thereby increasing the likelihood of confirmation bias and premature convergence. [6]

For example, in an intelligence analysis context an AI system may generate a prioritized list of potential threats based on historical patterns. Analysts may focus on the highest ranked option and allocate attention accordingly even when underlying data is incomplete or uncertain. Alternative assessments may receive less scrutiny because the AI output has implicitly defined what is most relevant.

Overtrust is reinforced by organizational incentives and performance narratives that prioritize speed and efficiency. When AI systems are considered remedial measures for overload or staffing constraints questioning their outputs can be perceived as inefficient or unnecessary. Responsibility for verification is diffused as outputs pass through multiple hands thereby obscuring who challenges validates or overrides AI-generated recommendations. [46]

This analysis does not assume that national security professionals are naive recipients of AI outputs. Many operators are acutely aware of AI system limitations, including hallucination risks and data provenance gaps. Therefore, the risk is not solely one of blind trust, rather it is one of structurally constrained judgment, a condition in which individuals who might otherwise contest AI outputs lack the institutional support, protected time, or organizational permission to do so. When organizations lack the structural capacity to verify and contest, the burden of managing AI reliability defaults to individual staff. When command guidance simultaneously frames cycle time as a decisive variable and organizations have yet to build the governance infrastructure to support meaningful verification, questioning AI outputs may remain formally permissible while becoming operationally costly in practice. In high-paced environments with strict speed mandates, this dynamic can lead to institutional metrics that register adoption and compliance, while operators at the execution point harbor unresolved skepticism about output reliability. This results in a gap that does not surface in aggregate productivity measures and must be addressed through specifically designed check-the-checker norms. Hence, the result can oscillate between institutional overtrust at the aggregate level and localized skepticism at the point of execution, a dynamic that governance frameworks that focus solely on automation bias are not designed to address.

Consequently, epistemic authority can shift incrementally from human judgment to machine-produced outputs, although formal decision authority remains nominally human. This shift is rarely explicit and is often difficult to detect until failure occurs. In the absence of clearly defined roles, training, and adjudication mechanisms, organizations risk normalizing reliance on AI outputs without clearly establishing the authority responsible for their correctness or consequences. Next, the study details the potential failure modes that may arise when AI adoption outpaces workplace readiness.

Failure Mode 1: Data Fidelity and Role Clarity Gaps

The first recurring failure mode emerges from gaps in workforce qualification, role definition, and data fidelity, resulting in a lack of organizational clarity regarding who to rely on or contest AI-enabled outputs. While national security organizations increasingly emphasize AI adoption, workforce frameworks do not consistently translate technical capability into operational role clarity. As a result, AI tools are often introduced into environments where personnel responsibilities for interpretation validation and escalation are not clearly defined.

From a workforce readiness perspective, this failure mode reflects a misalignment between skill development and authority allocation. Training individuals to operate AI tools without clearly defining responsibility for judgment and escalation creates conditions in which epistemic authority shifts through repeated use and practice. Without explicit workforce role clarity tied to data responsibility and adjudicative authority, AI adoption risks institutionalizing overtrust by default. When role-based qualifications and adjudication authority are lacking, AI outputs become operationalized through ambiguity.

This ambiguity is compounded by data fidelity challenges. AI systems often draw on heterogeneous data sources with varying levels of quality, provenance, and timeliness. When

personnel lack the training and explicit mandate to assess data limitations, AI outputs are potentially treated as more reliable than the underlying inputs justify, a dynamic consistent with documented automation bias and overreliance effects in decision-support systems.[4,5]

Public workforce frameworks, such as the Department of Defense Cyber Workforce Framework (DCWF) and the implementation of DoD Instruction 8140 establish baseline expectations for cyber and digital roles, including proficiency levels, training requirements, and certification pathways.[6] As AI-enabled systems increasingly shape decision processes, governance mechanisms that elucidate how authority is exercised in AI-supported environments complement current workforce structures. This alignment supports the development of formal, consistent practices that reinforce well-defined workforce roles.

Failure Mode 2: Governance Introduced After Acquisition

A second failure mode occurs when governance mechanisms are introduced after AI systems are acquired and deployed, rather than shaping system requirements and workforce preparation from the outset. In numerous cases, AI governance functions primarily as post-acquisition oversight, with limited influence on the design and applicability of systems.

The National Institute of Standards and Technology (NIST) Artificial Intelligence Risk Management Framework (AI RMF) emphasizes that effective AI governance must be integrated across the system lifecycle, including design, development, deployment, and use.[7] Similarly, the Department of Defense and the Chief Digital and Artificial Intelligence Office underscores embedding responsible AI principles early in acquisition and workforce preparation processes.[8] Despite these guidelines, governance measures are frequently introduced into systems only after the commencement of operational use.

Acquisition oversight bodies have repeatedly identified this pattern across technology programs. The Government Accountability Office (GAO) of defense acquisition programs note that delayed integration of governance and workforce readiness increases program risk and limits the effectiveness of downstream controls.[9] The introduction of governance after acquisition leads to organizations adapting supporting mechanisms around systems that are already shaping operational behavior.

This sequencing dynamic has direct workforce implications. Personnel are asked to use AI systems before their qualification and escalation standards are fully defined. Consequently, governance takes on a reactive posture, with workforce readiness being addressed later in the deployment process. Governance introduced after acquisition cannot fully compensate for workforce and accountability gaps that emerge during system design and fielding. When governance is attached as an afterthought after fielding, the organization cannot compensate for missing authority design, training gates, and verification roles. Consequently, governance capacity becomes the constraint on scaling AI beyond pilot use.

Failure Mode 3: Productivity Narratives and Verification Burden

The third failure mode emerges from productivity narratives that emphasize speed and efficiency while obscuring the downstream labor required to support AI-enabled outputs. AI adoption is frequently justified based on reduced workload and accelerated decision-making. These gains often depend on additional, and less visible, forms of human labor. Research on automation and socio-technical systems reveals that automated tools redistribute work. As systems increase in complexity, the labor associated with verification and coordination often increases, even when headline productivity metrics suggest efficiency gains.[10] This “hidden

labor" is essential to maintain system reliability but is frequently excluded from performance narratives used to justify adoption.

In national security contexts, verification burdens are particularly consequential. Personnel must reconcile AI outputs with existing intelligence, policy constraints, and command intent, often under time constraints. Prior research on automation bias and decision-support reliance indicates that when systems are framed as efficiency enhancers, users may reduce independent verification and critical review, further increasing organizational risk.[4,Error! Bookmark not defined.]

From a workforce perspective, this failure mode places additional strain on already constrained personnel while masking the true cost of AI integration. AI systems are frequently credited with dramatic time savings in terms of content generation; for example, these can produce analytic summaries or draft white papers in seconds as opposed to weeks. However, these apparent gains depend on downstream human labor to verify factual accuracy, assess currency, validate assumptions, and ensure alignment with doctrinal, policy, and formatting standards.

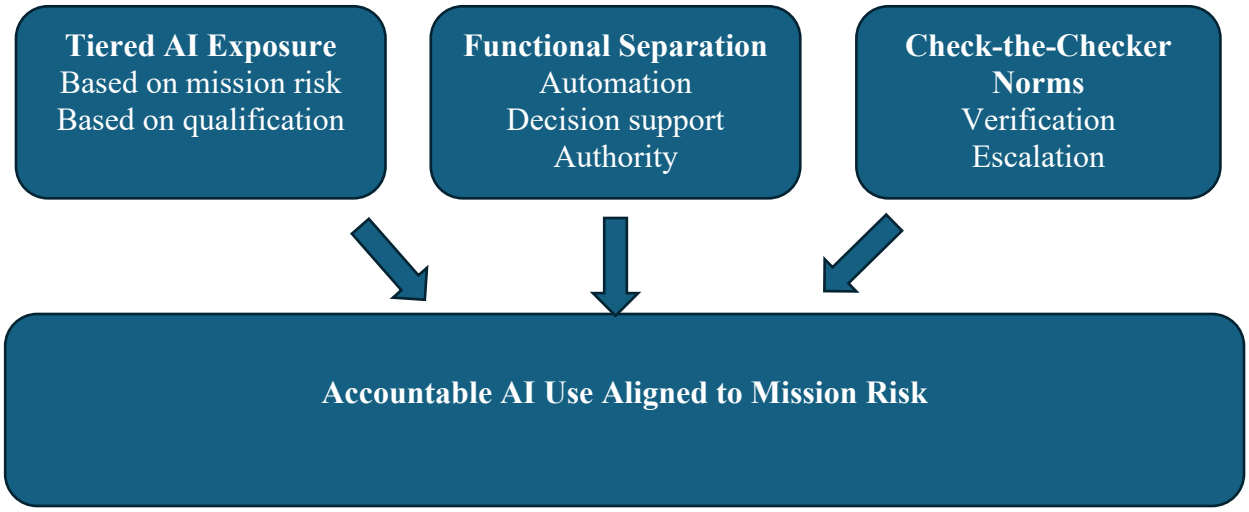
This verification and coordination are essential in national security contexts, yet it is rarely acknowledged in productivity metrics or workforce planning models. Therefore, productivity gains may appear as aggregate measures with accumulating verification burdens at the operational level, thereby increasing fatigue and reliance on informal judgment. Productivity narratives that ignore verification and coordination labor, risk undermining workforce readiness and creating hidden vulnerabilities in AI-enabled decision systems. Even when AI-enabled tools do reduce production time, organizations rarely articulate how the reclaimed time is reallocated, by whom, and toward which mission functions, leaving workforce expectations and

accountability structures implicit. Verification and coordination labor are governance functions performed by people. If these burdens are unplanned, workforce capacity becomes the limiting factor, and "speed gains" degrade into operational debt. The next section proposes a workforce governance model for national security AI adoption.

A Workforce-Centered Governance Model

Figure 1

Workforce-Centered AI Governance Model



Note. This model aligns AI capability with workforce readiness and authority structures through three reinforcing components. Tiered exposure gates AI use based on mission risk and qualification. Functional separation preserves the distinction between automation, decision support, and human authority. Check-the-checker norms institutionalize verification, escalation, and auditability. These elements convert AI capability into accountable action and reduce institutional overtrust.

The preceding sections demonstrated that near-term risks associated with AI adoption in national security organizations are attributed to the misalignment between technology, workforce

readiness, and governance structures, rather than to the autonomous system behavior. Addressing these risks requires reframing AI governance as a workforce and authority problem.

This section advances a workforce-centered governance model that aligns AI exposure, decision support, and accountability with mission risk and institutional readiness. The model proposes a workforce-centered governance model for national security AI adoption, considering workforce readiness and authority design as first-order controls. The model constitutes three mutually reinforcing components. First, it tiers AI exposure based on mission risk and demonstrated qualification, ensuring higher-consequence AI use is gated by readiness. Second, it demarcates automation, decision support, and authority so that AI assistance does not blur command accountability. Third, it institutionalizes check-the-checker norms through designated verification roles, escalation pathways, and auditability requirements. Together, these components convert AI capability into accountable action while reducing institutional overtrust.

Tiered AI Exposure Based on Workforce Readiness

Tiered AI exposure is the deliberate alignment of AI system access and use with workforce qualification, mission risk, and organizational readiness. This approach recognizes that different AI capabilities are associated with diverse operational, ethical, and strategic risk levels.

Hence, it reflects well-established principles to risk management in complex socio-technical systems, where access to higher-consequence capabilities is contingent on demonstrated knowledge, skills, accountability, and training. Tiering ensures that individuals and teams with well-defined responsibilities and appropriate preparation handle functions with greater operational impact. This approach creates marked distinctions between routine

automation, decision support that informs human judgment, and forms of use that shape or constrain high-consequence decisions.

From a mission assurance perspective, tiered exposure helps prevent premature reliance on AI outputs in contexts where workforce readiness, doctrine, and governance have not yet matured. It allows organizations to scale AI adoption deliberately by matching system capability to human capacity. In this model, AI exposure scales with workforce readiness and mission risk, thus, as safeguards develop, exposure levels can expand intentionally.

Separating Automation, Decision Support, and Authority

The model's second pillar is the explicit separation between automation, decision support, and authority. An ongoing conflict across these functions has directly contributed to overtrust in AI-enabled systems and to the erosion of accountability in operational settings. Automation involves systems that execute predefined actions upon activation, operating within defined rules and safeguards. Decision support systems, by contrast, generate recommendations, assessments, or predictions intended to inform human judgment. However, authority remains distinct from both as it is the human responsibility to make decisions, issue orders, and bear accountability for outcomes. Problems arise when these functions are implicitly blended in practice, even when they remain nominally distinct in policy or doctrine.

Human-in-the-loop and human-on-the-loop doctrines emphasize maintaining human judgment in decision-making processes involving automated systems particularly in military and other safety-critical contexts.[9] Human-in-the-loop refers to systems in which a human reviews and approves actions before execution. Human-on-the-loop refers to systems that operate under human supervision where a human can monitor intervene and override system behavior. Doctrine alone however is insufficient if workforce roles do not clearly define who may rely on

AI outputs who is responsible for validating them and who adjudicates disagreements between machine-generated recommendations and human judgment. Separating automation decision support and authority allows organizations to assign responsibility explicitly. This separation clarifies that authority cannot be delegated to AI systems even when automation or decision support scales extensively. This distinction preserves command accountability while allowing AI to augment human decision-making.

This separation encounters acute stress under hypercompressed operational timelines, where the physical time available for human adjudication is constrained, thus limiting the ability to exercise meaningful judgment before action is required. In conclusion, it is not that the doctrinal separation collapses rather it exposes where governance must operate: upstream, at the level of system design and mandate-setting. When an organization sets cycle time expectations that preclude verification, or deploys AI systems in contexts where human adjudication windows are measured in seconds, it has already made a consequential governance decision implicitly; without the accountability structures that explicit decision-making would require. The doctrinal separation between decision support and authority is operationally meaningful only if the conditions under which that separation is exercised are themselves governed. Therefore, governance of time pressure is governance of authority.

Institutionalizing "Check-the-Checker" Norms

The third pillar institutionalizes "check-the-checker" norms that formalize verification, escalation, and auditability for AI-enabled systems. Organizations should embed structured mechanisms to ensure that AI outputs are subject to appropriate scrutiny. In finance and other regulated industries, model risk management frameworks require independent validation, documented assumptions, and escalation protocols for decision-making models.[11] These internal

control models recognize that complex systems require oversight mechanisms that are organizational, not personal. Similar principles apply to AI in national security contexts. Check-the-checker norms include the following:

1) Defined verification roles responsible for assessing AI outputs against data quality, context, and mission intent.
2) Escalation pathways specifying when AI-human disagreements must be elevated.
3) Auditability and traceability requirements, documenting how AI outputs were generated, interpreted, and acted upon.

Embedding these norms reduces reliance on informal workarounds and mitigates the hidden labor problem identified earlier. It also reinforces institutional trust by making accountability visible and enforceable. Therefore, the key principle is as follows:

> *Trust in AI-enabled systems should be earned through institutional verification rather than inferred from system performance or presentation.*

Synthesis

Together, tiered exposure, functional separation, and check-the-checker norms form a workforce-centered governance model that aligns AI adoption with mission assurance. This model does not slow innovation; it disciplines it. By grounding AI integration in workforce readiness and authority structures, organizations can harness AI's benefits while preserving accountability in high-stakes environments. Table 1 presents the AI governance framework.

Table 1

Workforce-Centered AI Governance Framework

Use Category	Purpose	Who may use	Verification required	Escalation rule

Low-risk automation	Execute routine, predefined tasks where failure has limited mission impact (e.g., data formatting, routing, scheduling)	Personnel trained on the system; no special authority beyond role qualification	Spot checks; periodic audits of system performance and data inputs	Escalate only if system behavior deviates from expected parameters or produces repeated errors
Analytic decision support	Generate recommendations, summaries, or prioritizations to inform human judgment	Personnel with role-specific training and explicit authorization to use AI decision aids	Required human review; cross-checking against independent data sources when feasible	Escalate when AI output conflicts with human judgment, policy guidance, or mission context
Operational decision support (Moderate risk)	Support time-sensitive operational decisions that shape courses of action but do not execute them	Qualified personnel with advanced training and supervisory oversight	Mandatory verification by a second qualified individual or team; documented rationale for acceptance or rejection	Escalate all unresolved AI-human disagreements to a designated adjudicating authority
High-consequence decision shaping	Inform decisions with significant operational, strategic, or safety implications	Senior personnel with defined decision authority and documented AI training	Independent validation of assumptions, data sources, and model limitations; traceable documentation	Automatic escalation to command authority before action is taken
Prohibited or restricted uses	Applications where AI outputs would substitute for human judgment or blur command accountability	No authorized users absent explicit policy approval	Not applicable; use restricted or disallowed	Escalation triggered by attempted or unauthorized use

Note. Author's own work

Implementation Roadmap

The workforce-centered governance model outlined above is intentionally pragmatic. It is designed to be implemented incrementally within existing organizational structures. This section outlines a first-year implementation roadmap focusing on role clarity, training gates, escalation mechanisms, and acquisition integration. The objective is not to accomplish perfect AI governance in twelve months; rather, it is to establish durable foundations that prevent overtrust and preserve accountability as AI adoption accelerates.

First 12-Month Implementation Priorities

Define Workforce Roles for AI Use, Verification, and Adjudication

Organizations should explicitly define workforce roles related to AI-enabled systems. At minimum, this includes personnel authorized to use AI tools, personnel responsible for verifying AI outputs, and designated authorities empowered to adjudicate disagreements between AI recommendations and human judgment. These responsibilities should be documented in role descriptions, standard operating procedures, or mission directives. Clear role definition reduces the likelihood of epistemic authority shifting informally to AI outputs by default.

Introduce Training and Certification Gates Aligned to AI Exposure

Organizations should implement training and certification requirements tied to the tiered AI exposure model. Access to AI-enabled systems, particularly those used for analytic or operational decision support, should be contingent on demonstrated readiness. Training should address system operation, data limitations, uncertainty, failure modes, and escalation expectations. This approach aligns with phased adoption models in safety-critical domains where access expands as readiness matures.[12]

Formalize Escalation Rules for AI-Human Disagreement

Escalation thresholds should be clearly defined and communicated. Personnel should not be left to infer when disagreements between AI outputs and human judgment require elevation. Conditions such as conflicting intelligence assessments, policy ambiguity, or high-consequence decisions should trigger predefined escalation pathways. Clear escalation rules reduce cognitive burden and reinforce that questioning AI outputs is an expected responsibility.

Account for Verification and Coordination Labor in Workforce Planning

Verification and coordination are governance functions performed by people and must be explicitly accounted for in workforce planning. Organizations should establish a baseline assessment of verification time relative to AI-assisted production time across use categories. Where verification time approaches or exceeds AI-enabled time savings, the productivity rationale for AI adoption should be reevaluated. Even rough estimates are sufficient to determine whether AI adoption is generating real capacity or redistributing hidden labor through fatigue, shortcuts, or degraded accountability.

Integrate Workforce and Governance Requirements into Acquisition Processes

AI system requirements should incorporate workforce and governance considerations from the outset. In addition to technical performance, requirements should specify training needs, verification roles, documentation standards, and auditability expectations. Embedding governance into acquisition prevents the need to retrofit oversight mechanisms after deployment and aligns with lifecycle governance principles outlined in frameworks such as the NIST AI Risk Management Framework and Department of Defense AI guidance.[8] [9].

Conclusion

AI is often framed as a technical or strategic breakthrough that national security organizations must adopt rapidly to remain competitive. This study contends that such framing is inadequate. The primary risk associated with near-term AI adoption is not autonomous system behavior or machine agency, but institutional overtrust arising from misaligned workforce readiness, unclear authority boundaries, and governance structures that lag deployment. When AI systems generate outputs in answers-like forms, they reshape how knowledge, judgment, and authority circulate inside organizations, often without explicit acknowledgment or design.

Across defense and national security contexts, the pressure to accelerate has elevated productivity, speed, and advantage as dominant narratives for AI adoption. Yet without clear role definitions, training gates, escalation pathways, and verification norms, these pressures can erode accountability and obscure responsibility. AI is not a replacement for human decision-makers to undermine mission assurance. However, its use can be implemented incrementally by narrowing deliberation, anchoring judgment, and diffusing adjudicative authority across systems and personnel unprepared to manage its epistemic weight.

Reframing AI adoption as a workforce and governance challenge provides a more durable path forward. By aligning AI exposure with readiness, separating automation from decision support and authority, and institutionalizing “check-the-checker” norms, organizations can integrate AI in ways that can strengthen command accountability. These measures do not slow innovation; rather, they discipline it, ensuring that the acceleration gained through AI is not at the expense of judgment, trust, or mission assurance.

Ultimately, the prime concern of national security leaders is not whether to adopt AI; but how to adopt it responsibly under uncertain conditions and strategic competition. In conclusion,

innovation that outpaces governance can hollow out the very decision systems it seeks to enhance. However, innovation aligned with workforce readiness and authority provides a sustainable foundation for operational advantage in contested and high-stakes environments.

Endnotes

[1] White House. (2025a, July). *America's AI action plan*. https://www.whitehouse.gov/wp-content/uploads/2025/07/Americas-AI-Action-Plan.pdf

[2] White House. (2025b, July). *White House unveils America's AI action plan*. https://www.whitehouse.gov/articles/2025/07/white-house-unveils-americas-ai-action-plan/

[3] Secretary of War. (2026, January 9). *Artificial intelligence strategy for the department of war: Accelerating America's military AI dominance* [Memo]. https://media.defense.gov/2026/Jan/12/2003855671/-1/-1/0/artificial-intelligence-strategy-for-the-department-of-war.PDF

[4] Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

[5] Mosier, K. L., Skitka, L. J., Heers, S., & Burdick, M. (1996). Automation bias: Decision making and performance in high-tech cockpits. The *International Journal of Aviation Psychology*, *6*(1), 47–63. https://doi.org/10.1207/s15327108ijap0801_3

[6] Department of Defense. (2023). *DoD instruction 8140.03: Cyberspace workforce qualification and management program*. https://dodcio.defense.gov/Portals/0/Documents/Library/DoDM-8140-03.pdf

[7] National Institute of Standards and Technology. (2023). *Artificial intelligence risk management framework (AI RMF 1.0)*. https://www.nist.gov/itl/ai-risk-management-framework

[8] Department of Defense. (2024). *Responsible artificial intelligence strategy and implementation pathway* (Report No.). https://media.defense.gov/2024/Oct/26/2003571790/-1/-1/0/2024-06-RAI-STRATEGY-IMPLEMENTATION-PATHWAY.PDF

[9] Government Accountability Office. (2022). *Defense acquisitions: DOD should take additional steps to assess the performance of its major weapon programs* (Report No. GAO-22-104689). https://www.gao.gov/products/gao-22-104687

[10] Suchman, L. (2007). *Human-machine reconfigurations: Plans and situated actions* (2nd ed.). Cambridge University Press. https://doi.org/10.0000/0000

[11] Board of Governors of the Federal Reserve System. (2011). *Supervisory guidance on model risk management* (Report No. SR 11-7). https://www.federalreserve.gov/supervisionreg/srletters/sr1107.htm

Escalation Protocols for Autonomous Naval Vessels:
A Legal-Technical Framework for Safe and Credible Maritime Autonomy

LCDR Jordan J. Foley, JAGC, U.S. Navy

MIT–Harvard Technology & National Security Conference (April 3–4, 2026)

Author Note

The views expressed in this article are those of the author alone and do not reflect the official policy or position of the Department of Defense, the Department of the Navy, or any other agency of the U.S. Government. This publication is intended for academic and policy discussion purposes and does not constitute legal advice or a formal endorsement of any particular course of action.

Author Bio

Jordan J. Foley is a Lieutenant Commander in the U.S. Navy Judge Advocate General's Corps and an experienced national security and maritime law attorney. He currently advises senior Navy leadership on admiralty, tort, and international maritime law while managing complex litigation in coordination with the Department of Justice and Department of Defense. Foley holds degrees from the Georgetown University Law Center, Massachusetts Institute of Technology, Naval War College, and the United States Naval Academy, and his scholarship focuses on maritime law, national security, and emerging technologies in defense.

Abstract

Autonomous and remotely supervised surface vessels are moving from pilots to operational deployment across defense and commercial shipping. U.S. Navy unmanned surface vessels (USVs)—including mine countermeasures platforms—and commercial Maritime Autonomous Surface Ships (MASS) operate in congested waterways alongside human mariners yet remain governed by legal regimes built around human signaling and post-incident accountability.[1][2] This paper argues that maritime autonomy faces a governance vulnerability: systems may navigate safely but remain legally brittle if their conduct cannot be rendered intelligible under the standards used to assign responsibility after incidents.

Maritime compliance is evaluated through reconstruction. The COLREGs apply to "all vessels," require standardized lights and sound signals, and permit supplemental warnings under Rule 36 only if they cannot be mistaken for authorized signals.[3] UNCLOS reinforces compliance as sovereign responsibility by requiring flag states to ensure conformity with generally accepted international safety regulations.[4] These rules create an evidentiary burden autonomy often cannot meet. Collision doctrine underscores the stakes: statutory navigation violations can trigger presumptions of fault that must be rebutted by proof.[5] In coercive or ambiguous encounters, the burden intensifies because escalation governance depends on demonstrable sequencing and restraint rather than after-action narrative.[6]

Existing maritime systems generate abundant data but rarely preserve a unified, time-synchronized record of what an autonomous vessel perceived, what it signaled, how it maneuvered, and whether human-on-the-loop judgment boundaries were maintained.[7] The paper proposes a maritime assurance operating layer as the enabling infrastructure for lawful autonomy at scale. Non-controlling and platform-agnostic, it does not steer the vessel or execute coercive actions; it fuses heterogeneous sensor streams and operational logs into an audit-grade timeline that captures signaling and escalation sequencing, including Rule 36 warnings and human

intervention pathways.

By making compliance *reconstructable* and restraint provable, assurance infrastructure reduces legal exposure, narrows narrative ambiguity in contested waters, and supports coalition and commercial adoption through portable accountability outputs. The core conclusion is that maritime autonomy will scale only if it becomes legible: evidence-grade assurance is a condition of credible, stabilizing autonomy at sea.

I. Introduction: Autonomy's Legibility Gap

Autonomous and remotely supervised surface vessels are transitioning from pilots to operational deployment across defense and commercial maritime sectors. In this paper, military unmanned surface vessels are treated as unmanned surface vessels (USVs), including the U.S. Navy's Mine Countermeasures Unmanned Surface Vehicle program and other government-operated platforms used for mine warfare, reconnaissance, and force protection.[8] The Navy's mine countermeasures integration offers a timely case: the first Mine Countermeasures Mission Package embarked on USS Canberra (LCS 30) in April 2024, reflecting movement from experimentation toward fielded capability,[9] with continued reporting as deployment timelines mature.[10] In parallel, regulators and industry use Maritime Autonomous Surface Ships (MASS) to describe commercial autonomy across varying degrees of remote supervision and automated control.[11] USVs and MASS differ in mission and legal posture—including sovereign immunity and defense rules of engagement—but both operate in congested waterways alongside human mariners and must comply with legal regimes built around human bridge teams and human signaling practices.[12] That shared operating context produces a common governance vulnerability: autonomy may be technically feasible, yet legally fragile when its behavior cannot be rendered intelligible under the standards used to assign responsibility after incidents.

Maritime governance is fundamentally post-incident. The COLREGs apply broadly "to all vessels"[13] and define "vessel" expansively.[14] The rules are widely treated as reflecting customary international law and baseline expectations for safe navigation.[15] But key obligations—lookout, collision-risk assessment, and early maneuver—presume the "ordinary practice of seamen," pushing autonomy governance toward functional-performance interpretations rather than literal crew-based assumptions.[16] The IMO's scoping work reflects the same tension: MASS is expected to comply with existing instruments even as some provisions require interpretation or adaptation for remote and autonomous modes.[17]

For autonomy, the distinction between safe navigation and defensible navigation is decisive.

Liability and accountability turn on what can be demonstrated through credible evidence—not what designers intended. Collision doctrine illustrates the point: statutory navigation violations can trigger presumptions that must be rebutted by proof.[18] As autonomy scales, this proof burden becomes harder because consequential decisions are made by software systems whose internal states are not inherently visible to investigators, insurers, regulators, or commanders.

The governance problem intensifies in coercive or ambiguous encounters. Even outside armed conflict, vessels may face unsafe approaches, harassment, or interference requiring warnings and graduated deterrence under operational policy.[19] International navigation law anticipates attention-getting warnings beyond standard signals: Rule 36 permits additional light or sound signals so long as they cannot be mistaken for authorized COLREG signals.[20] But permissibility is not enough. If an encounter becomes disputed, the operator must be able to prove what warnings were issued, when, and under what conditions—especially given flag-state duties to ensure compliance with generally accepted safety regulations.[21]

This paper argues that maritime autonomy requires an assurance operating layer designed to produce legally relevant legibility. The layer does not steer the vessel or replace autonomy stacks. It functions as accountability infrastructure: synchronizing heterogeneous sensor streams and operational logs into a *reconstructable* timeline capable of showing what the vessel perceived, what it signaled, how it maneuvered, and what human-on-the-loop interventions occurred—so conduct can be reviewed against COLREG duties and operational governance constraints.[22] In the TechNatSec frame, the claim is practical and forward-looking: if autonomy is "innovating on the frontlines," then evidence-grade legibility is the enabling infrastructure that makes deployment governable, interoperable, and strategically stable.

II. Escalation of Force at Sea: Legal Requirements vs. Technical Reality

Escalation-of-force at sea functions less as a tactical ladder than as a governance mechanism designed to reduce miscalculation and ensure coercive actions remain lawful, necessary, and proportionate under domestic policy and international law.[23] Operationally, it embeds sequencing and restraint: warnings precede compulsion, reversible measures precede irreversible ones, and

lethal force—if ever authorized—remains tightly bounded by legal necessity and command oversight.[24]

These constraints are especially salient at sea, where encounters are fast-moving and ambiguous and misinterpretation can escalate routine interactions into strategic incidents.[25] However, escalation doctrine assumes human perception, judgment, and documentation. Autonomous and remotely supervised vessels disrupt that model. The central challenge is not whether unmanned systems can execute graduated warnings and deterrence, but whether they can do so in a way that remains legally reviewable and strategically credible. Compliance cannot be inferred from design intent; it must be demonstrated through a reconstruction-grade record of what occurred and why.[26]

A. EOF Doctrine and Sequential Restraint

Escalation practice in maritime operations is structured around sequential restraint. U.S. doctrine and operational policy require that force be employed consistently with mission needs, escalation control, and legal constraints, with actions governed by command authority rather than improvisation.[27] The Standing Rules of Engagement and Standing Rules for the Use of Force provide the baseline architecture for graduated response options, reinforcing that warnings and deterrence measures should be applied in a controlled sequence that supports proportionality and accountability.[28]

This sequencing logic mirrors international constraints on coercive action at sea. UNCLOS prohibits the threat or use of force inconsistent with the U.N. Charter, underscoring that navigation and maritime operations cannot be used as vehicles for unlawful coercion.[29] At the same time, international law preserves limited pathways for lawful force, including self-defense and certain enforcement actions.[30] International jurisprudence has also emphasized necessity and proportionality in maritime enforcement, including scrutiny of excessive coercive measures against vessels.[31]

These principles are operationally consequential for unmanned systems. USVs conducting sensitive missions—such as mine countermeasures—are likely to encounter ambiguous

approaches and harassment where warning and deterrence options are necessary but must remain tightly governed.[32] Commercial MASS platforms face similar close-approach risks and interference scenarios while remaining subject to COLREG sound-signal and warning mechanisms.[33] In both contexts, escalation is not simply whether warning tools exist; it is whether the vessel can demonstrate that warnings were context-appropriate, non-misleading, and properly sequenced under safety and proportionality constraints.[34]

These evidentiary demands intensify in gray-zone competition, where ramming, water cannons, and aggressive maneuvering may be used to create ambiguity while avoiding conventional thresholds.[35] In such environments, credibility depends not only on restraint in practice, but on the ability to prove that restraint through a coherent record of escalation steps and decision timing.[36]

B. DoDD 3000.09 and the Demand for Reviewability

The doctrinal demand for sequencing becomes a technical governance requirement under modern autonomy policy. DoD Directive 3000.09 requires autonomous weapons systems to be used with "appropriate levels of human judgment," reinforcing that autonomy does not displace accountability.[37] While the directive is often framed around kinetic targeting, its operational significance for maritime autonomy is broader: when autonomy enables actions with coercive effect, the system must remain governed by auditable human authority and policy constraints.[38]

This requirement is directly relevant to maritime signaling and deterrence systems, where many escalation measures are technically "non-lethal" but operationally coercive. These include acoustic hails, long-range warning devices, and entanglement systems intended to disable propulsion or enforce standoff distance.[39] Such tools may be permissible, but permissibility is contextual—dependent on warnings, proportionality, and authorized control.[40] Even capabilities treated as "distraction devices" remain governed by acquisition and employment frameworks that assume defined conditions, training, and oversight.[41]

Autonomous systems stress these assumptions because they compress decision timelines and

distribute perception across sensors rather than human observation. Traditional accountability often relies on crew recollection, bridge logs, and testimony, all of which can be incomplete even for crewed vessels. The challenge is amplified for USVs and MASS: remote operators may lack the situational cues of an onboard watch team, and autonomy stacks may classify encounters in ways that are difficult to translate into legally meaningful categories without structured interpretation.[42]

Accordingly, the "appropriate human judgment" requirement must be operationalized through demonstrable human-on-the-loop boundaries and traceable event reconstruction.[43] At minimum, the system must preserve evidence of what it perceived, what it signaled, what escalation option was selected, whether and when a human authorized that step, and how quickly the sequence unfolded. Without this record, compliance cannot be meaningfully assessed and accountability collapses into unsupported assertions.[44]

C. The Failure of After-Action Narratives Without Evidence

In maritime law and operations, post-incident narratives are not merely descriptive—they determine responsibility. Investigators, insurers, commanders, and courts evaluate compliance using standardized rules and presumptions that reward objective proof rather than subjective explanation. U.S. collision doctrine illustrates this rigor: a vessel violating a statutory navigation rule is presumed at fault unless it proves the violation could not have contributed to the casualty.[45] Courts have also held that reasonable seamanship can require adoption of available safety technologies, underscoring that compliance expectations evolve with operational reality.[46] For autonomy, the implication is direct: credibility depends on demonstrable compliance, not post hoc intent.

Autonomous systems strain this evidentiary model because they frequently fail to generate a unified record that can withstand scrutiny. Traditional vessel logs were built for human crews and rarely preserve synchronized, multi-sensor reconstruction. At the same time, proprietary autonomy logs are often non-standard, difficult for third parties to interpret, and poorly suited for adversarial review. The predictable outcome is that even safe autonomous behavior can become

contested when the system cannot coherently reconstruct what occurred, delaying adjudication and increasing operational and legal risk.

This evidentiary gap is not theoretical. Maritime signaling and collision-avoidance regimes depend on recognizability and clarity, because ambiguity can itself constitute negligence or breach of the duty of care.[47] Flag states must ensure vessels conform to generally accepted international safety regulations, including COLREG requirements.[48] The COLREGs further require vessels to take all necessary precautions to avoid immediate danger, including measures beyond rigid rule compliance.[49] For autonomous operations, these standards create a governance requirement: systems must not only act safely, but produce evidence sufficient to prove safe and compliant conduct under post-incident review.

The need for reconstruction-grade evidence becomes urgent in gray-zone maritime environments where ambiguity is weaponized. Coercive practices such as ramming and water cannon use exploit the gap between routine navigation and overt armed force while shaping public narratives in real time.[50] In these conditions, "what happened" is contested immediately. If a USV or MASS platform cannot produce a defensible record of warnings, signaling compliance, and proportional restraint, operator credibility can degrade regardless of actual behavior.

The conclusion is foundational: autonomy cannot scale on performance claims alone. Compliance cannot be inferred from design intent or vendor assurance; it must be demonstrated through an evidentiary record that survives legal scrutiny and strategic contestation. For this reason, escalation governance in unmanned maritime systems depends on technical infrastructure that treats post-incident legibility as a primary requirement—not an accidental byproduct of navigation systems. Subsequent sections build that case by specifying where current maritime architectures fail and what an assurance operating system must provide to make graduated deterrence credible, reviewable, and defensible for both defense and commercial autonomy.

III. Visual Signaling, COLREGs, and the Functional Equivalence Problem

Autonomous maritime operations face a simple legal constraint with difficult implementation:

the COLREGs apply to "all vessels," yet were written for human mariners.[51] Because the rules are widely treated as customary international law, they carry binding practical force across jurisdictions and operating contexts.[52] The dominant governance response is functional equivalence: autonomy must meet the performance and signaling expectations the rules demand even without a human bridge team.[53] But functional equivalence is not merely theoretical. It is a proof burden. If an incident becomes contested, the operator must be able to demonstrate compliance in forms third parties can evaluate under established accountability logic.[54]

Rule 36 is the most direct example. It permits supplemental light or sound signals to attract attention only if those signals cannot be mistaken for authorized COLREG signals.[55] For autonomy, that means legality turns on more than whether a warning device exists. The system must be able to show what signal was emitted, that it was non-confusable, and that it was issued in a timely manner relative to encounter dynamics. Where a platform's construction or purpose makes strict compliance impracticable, Rule 1(e) provides a "closest possible compliance" channel through flag-state determination.[56] That, too, creates an evidentiary requirement: the operator must be able to demonstrate the applicable compliance posture and the basis for it.

The consequence is operational and strategic. In collisions and close-approach events, signaling disputes are often dispositive because maritime law heavily weights recognizability and clarity.[57] Flag states bear obligations to ensure vessels conform to generally accepted safety regulations.[58] For autonomy, signaling without logging is therefore legally insufficient: the ability to prove what was displayed and why becomes part of compliance itself.[59]

IV. Why Existing Maritime Systems Fail

Autonomy's legal durability depends on *reconstructability*: a unified record showing what the vessel perceived, what it communicated, how it maneuvered, and how escalation-related steps were sequenced.[60] The problem is structural. Existing maritime systems generate extensive data, but they were built for real-time navigation by human bridge teams, not for audit-grade reconstruction of autonomous conduct.

AIS, radar, ECDIS, and VDRs each contribute partial visibility, but none provides an end-to-end, time-synchronized account of perception, signaling, and decision pathways. AIS can approximate position and course but cannot establish COLREG compliance with lookout, risk assessment, or early maneuver—and is unreliable in adversarial settings where transmissions can be degraded or manipulated.[61] Radar preserves tracks but not communications or warning actions; ECDIS shows where the vessel went but rarely why; and VDRs vary by implementation and typically do not capture distributed autonomy decision outputs or remote human-on-the-loop interactions in a form suitable for policy review.[62] Under regimes that impose strict proof burdens after incidents, these gaps are not academic—they are liability and credibility vulnerabilities.[63]

Autonomy stacks deepen the gap. "Stacked" architectures distribute legally salient decisions across proprietary perception, classification, and control modules that are not standardized and are often not interpretable in legal terms.[64] Yet the legal questions autonomy must answer are inherently reconstructive: whether warnings were issued under Rule 36 and were non-confusable, whether any alternative compliance posture applied, and whether escalation-related steps remained reviewable under policy constraints.[65] Without an auditable record tying sensor perception to signaling and action sequencing, investigations default to competing narratives—exactly the instability that gray-zone actors exploit.[66]

V. The Maritime Assurance Operating Layer: Technical Architecture

Maritime autonomy operates under a central constraint: compliance must be demonstrable, not presumed. Executing collision-avoidance maneuvers or broadcasting standard signals is necessary but insufficient. Legal durability depends on whether an autonomous vessel can produce a coherent record showing compliance with the COLREG signaling regime, satisfaction of flag-state safety obligations, and execution of escalation steps consistent with policy constraints and lawful restraint.[67]

This section advances the core technical proposition: autonomous maritime operations require a separate maritime assurance operating layer—an independent system that does not steer the

vessel, select maneuvers, or control weapons, but renders behavior legible, reviewable, and *reconstructable* after the fact. It is infrastructure rather than a software feature, combining time-synchronized sensor capture, operational logging, and interpretability mechanisms suitable for mixed traffic, remote supervision, COLREG signaling requirements, and escalation governance boundaries.

This assurance layer is necessary because today's maritime evidence ecosystem cannot reliably prove lawful conduct. Under *The Pennsylvania* rule, violation of a statutory navigation requirement triggers a presumption of fault that must be rebutted through evidence.[68] Flag states must ensure compliance with generally accepted international safety regulations, including COLREG obligations governing signaling and avoidance behavior.[69] Operational policy similarly requires sequential restraint and controlled authority in escalation, supported by reviewable chains of responsibility.[70] In autonomy-enabled operations, these legal and policy requirements translate directly into engineering requirements

A. Design Requirements Derived from Law

The maritime assurance operating layer begins with a governance premise: autonomous vessels must be evaluated not only by outcomes, but by whether they performed the legally required steps that enable safe and predictable navigation. The COLREGs apply broadly to "all vessels," and define "vessel" expansively without limiting application based on crewing model, indicating that unmanned and remotely supervised craft remain within scope.[71] Functional equivalence frameworks therefore require autonomous vessels to meet the performance and signaling expectations traditionally fulfilled by human bridge teams.[72]

Critically, functional equivalence is an evidentiary burden. Autonomous vessels must be able to demonstrate that they took required precautions to avoid collision and attract attention, including compliance with standardized lights, shapes, and sound signals, and the appropriate use of supplemental warnings under Rule 36 when necessary.[73] Rule 2 reinforces that safe navigation is assessed holistically: vessels must take "all necessary precautions" to avoid immediate danger even beyond strict rule text.[74] For autonomy, these obligations require post-incident proof—not

only that behavior was safe, but what precautions were taken and why the circumstances required them.

UNCLOS anchors this compliance obligation at the flag-state level. States must maintain effective control and ensure conformity with generally accepted international safety regulations, including collision-prevention requirements.[75] In autonomy settings, "effective control" cannot be reduced to remote supervision alone; it must include the ability to show that vessel conduct was constrained by safety rules and accountable decision pathways.

Escalation governance adds further constraints. U.S. operational frameworks require sequential restraint under the Standing Rules of Engagement and Standing Rules for the Use of Force.[76] International law similarly prohibits threats or uses of force inconsistent with the U.N. Charter.[77] International jurisprudence has also assessed maritime coercion under necessity and proportionality standards, including where enforcement actions exceed lawful bounds.[78] For autonomy-enabled vessels capable of issuing warnings, deploying non-lethal measures, or taking deterrent actions, these principles impose a technical requirement: the system must produce a record that demonstrates sequencing, proportionality context, and oversight boundaries.

From these constraints, the assurance operating layer must meet five design requirements:

1. Non-controlling: It must not direct navigation or operational decisions in real time, reducing the risk it is characterized as a weapons controller or maneuvering authority.[79]
2. Sensor-agnostic: It must integrate heterogeneous sensors and vessel systems without requiring adoption of a single proprietary autonomy stack.[80]
3. Legally aware: It must represent behavior in rule-relevant terms such as signaling compliance, encounter type, proximity, and warning sequence.[81]
4. Time-synchronized: It must align evidence across sensors, logs, and human inputs into a coherent timeline suitable for litigation-grade reconstruction.[82]
5. Audit-grade: It must preserve immutable, reviewable records adequate for investigations, underwriting, and policy review under flag-state and operational governance obligations.[83]

This architecture functions as infrastructure: it enables safety and accountability independent of mission profile or vendor implementation. It is the difference between autonomy that is merely operational and autonomy that is legally durable.

B. Sensor Fusion and Temporal Alignment

The most immediate technical obstacle to lawful autonomy at scale is fragmentation. Maritime evidence sources—AIS, radar, ECDIS records, VDR outputs, camera feeds, and autonomy logs—typically exist as parallel streams that are rarely synchronized into a unified event record. As a result, post-incident reconstruction often cannot answer basic questions with precision: what the vessel perceived, when it perceived it, and what it did in response. Historically, human bridge teams supplied narrative context through testimony and log entries. Autonomy reduces the availability and reliability of that scaffolding, increasing dependence on machine-generated reconstruction.[84]

A maritime assurance operating layer addresses this gap by treating time-synchronization as a core system requirement. Both navigation and escalation compliance turn on sequencing. COLREG analysis often depends on whether risk was assessed correctly and whether maneuvering and signaling occurred early enough to reduce danger.[85] Escalation compliance is even more timing-dependent: warnings must precede compulsion, reversible measures must precede irreversible ones, and high-consequence actions require preserved human authority boundaries.[86]

Accordingly, the assurance layer should fuse heterogeneous sources into a single timeline that can reconstruct both encounter dynamics and system response. At minimum, this includes radar track histories; AIS transmissions and discrepancies; ECDIS navigation states and route history; available visible/infrared video; audio records of hails or warning devices; and autonomy outputs showing perception classifications and recommended responses.[87] The objective is interoperability and interpretability—not replacement of underlying sensors or autonomy stacks. This approach aligns with the MASS regulatory trajectory. IMO and national authorities are

evaluating how existing requirements can be satisfied through alternative means in remote and automated modes, implicitly demanding architectures capable of demonstrating compliance across distributed systems.[88]

Temporal alignment must also be verifiable. A unified record is only as reliable as its synchronization. If timestamps drift or data lacks consistent time sources, sequencing becomes contestable and reconstruction fails under scrutiny. In collision doctrine, ambiguity frequently cuts against a vessel that cannot produce clear evidence to rebut presumptions of fault.[89] In autonomy-enabled operations, that burden is amplified because decision-making is not directly observable without synchronized records.

C. EOF Sequencing and Warning Logs

A second critical function of the assurance layer is producing a standardized sequencing log for warnings, signaling, and deterrence actions. This is where legal compliance translates directly into technical requirements. COLREG Rule 36 permits supplemental light or sound signals to attract attention, but only if those signals cannot be confused with authorized signals elsewhere in the rules.[90] Rule 1(e) permits alternative compliance pathways for vessels of special construction or purpose when determined by the flag state.[91] Operational policy similarly requires warnings and graduated measures to reflect restraint and accountability.[92]

Accordingly, the assurance layer should log signaling and deterrence behavior in a format that supports review. At minimum, the record must capture the signal type (visual, acoustic, VHF, or other attention-getting method), whether it was COLREG-authorized or a Rule 36 supplemental warning, and any duration or intensity parameters where relevant. It must also preserve the triggering conditions (e.g., range, closing speed, CPA threshold) and the decision pathway—whether the action was system-initiated, system-recommended, or executed only after human confirmation.[93]

This capability does not create new law; it operationalizes existing law. Without logging, signaling becomes legally contestable, and inability to demonstrate compliance can translate into

inability to rebut liability under established presumptions.[94] In gray-zone environments where coercive tactics are used to provoke incidents and contest legitimacy, an auditable warning record also becomes strategically significant.[95]

The assurance layer must further support review of non-lethal measures. Maritime interdiction tools such as entanglement systems have been recognized as non-lethal techniques when used with safeguards including warning and clear identification of noncompliant vessels.[96] While legality depends on context, the governance requirement remains stable: if coercive measures are employed, the system must preserve evidence supporting warning, necessity, and proportionality. International jurisprudence has evaluated maritime coercion under these standards, including in *M/V Saiga* and *Guyana v. Suriname*.[97]

By capturing warning sequences alongside encounter context, the assurance layer converts escalation governance from a subjective narrative into an auditable record.

D. Human-on-the-Loop Capture Without Command Authority

The final design requirement is preserving human accountability without turning the assurance layer into a command-and-control system. DoDD 3000.09 requires autonomous weapons systems to be used with "appropriate levels of human judgment," establishing a baseline principle of responsible autonomy governance.[98] In maritime operations, the practical extension is that escalation-related actions—especially those with coercive effects—must remain reviewable against defined human intervention boundaries.

The assurance layer supports this requirement by capturing human-on-the-loop inputs as evidence rather than as real-time command logic. At minimum, it should record whether an operator received an alert, what information was presented, what response options were available, whether the operator approved, denied, or deferred action, and the elapsed time between recommendation and execution.[99] This is a governance function, not a maneuvering function. The assurance layer does not decide whether to escalate; it preserves the record needed to evaluate restraint, oversight, and decision authority under review.

This distinction preserves the layer's non-controlling character. If the assurance layer were responsible for navigation or coercive decisions, it would become operationally entangled with the autonomy stack and undermine its role as an independent accountability system. By remaining non-controlling, the layer can be implemented across platforms—including USVs and MASS—without requiring redesign of core control logic.

In sum, autonomy at scale requires more than safe performance; it requires provable compliance. The maritime assurance operating layer enables audit-grade reconstruction of signaling and escalation behavior by integrating time-synchronized evidence, standardized sequencing logs, and human oversight capture. By translating COLREG compliance, flag-state obligations, and autonomy governance requirements into system architecture, it offers a practical path toward lawful and credible maritime autonomy.[100]

VI. Governance, Risk, and Strategic Stability Implications

Autonomous maritime platforms will not be adopted at scale simply because they can navigate efficiently. Adoption depends on whether autonomy can be governed credibly under legal regimes that regulate navigation, constrain coercion, and assign responsibility after incidents. In practice, credibility is tested in moments of friction—unsafe approaches, contested encounters, collisions, interference, and ambiguous uses of force below kinetic thresholds. In those conditions, autonomy is judged not only by immediate safety outcomes, but by whether conduct can be defended as lawful and restrained through an evidentiary record capable of withstanding scrutiny.

An assurance operating layer is therefore more than a technical fix for evidence fragmentation; it functions as a strategic stabilizer. By making restraint provable, it reduces escalation risk, narrows narrative ambiguity in gray-zone encounters, and enables coalition interoperability through shared accountability standards across divergent autonomy stacks.[101]

A. Legal Defensibility in Contested Waters

The most immediate governance value of assurance infrastructure is that it converts lawful conduct from an assertion into a demonstrable fact. Maritime legal frameworks are enforced through reconstruction. COLREG compliance turns on whether a vessel took appropriate precautions, signaled clearly, and maneuvered safely in context.[102] UNCLOS reinforces this compliance posture by requiring flag states to ensure vessels conform to generally accepted international regulations.[103] U.S. operational doctrine similarly expects escalation steps to be reviewable under structured rules governing warning, deterrence, and restraint.[104]

For autonomous vessels, these regimes create a structural risk: safe behavior is not sufficient if the vessel cannot prove what it did and why. The evidentiary discipline of collision law illustrates the point. Under *The Pennsylvania* rule, violation of a statutory navigation requirement triggers a presumption of fault, shifting the burden to prove the breach could not have contributed to the incident.[105] Courts have also emphasized that signal clarity and recognizability are central to the duty of care, meaning ambiguity in signaling can itself be negligence.[106] In contested operating environments—where adversaries may seek to provoke incidents and contest responsibility—these doctrines produce a governance trap: the party that cannot produce evidence is disadvantaged regardless of whether it acted correctly.

Assurance infrastructure changes this dynamic by generating an audit-grade record that supports legal defensibility. A synchronized timeline can demonstrate whether an autonomous vessel complied with signaling requirements, whether it used Rule 36 supplemental warnings without creating confusion, and whether maneuver timing aligned with collision-avoidance obligations.[107] In escalation contexts, it can demonstrate sequential restraint—warnings before compulsion, reversible measures before irreversible ones—and preserve evidence of human-on-the-loop intervention boundaries for higher-consequence actions.[108] Operationally, this reduces legal uncertainty and enables commanders and policymakers to field autonomy without accepting unmanaged accountability risk.

B. Sovereign Immunity and Narrative Warfare

Assurance infrastructure also shapes strategic stability by strengthening sovereign immunity posture and reducing vulnerability to narrative manipulation. Sovereign immunity is central to military vessel governance. UNCLOS recognizes that warships and other sovereign immune vessels are not subject to foreign jurisdiction in the same way as commercial ships.[109] U.S. legal doctrine similarly recognizes sovereign immunity for government vessels.[110] Yet even where immunity is legally robust, legitimacy is politically contestable. States can still face strategic costs if maritime actions are portrayed as unsafe, unlawful, or escalatory.

Gray-zone dynamics illustrate why narrative competition matters. Contemporary maritime coercion frequently relies on ramming, water cannons, and aggressive maneuvering to impose pressure while remaining below conventional thresholds of armed conflict.[111] The South China Sea arbitration criticized China's water cannon use as dangerous and disproportionate against Philippine vessels, demonstrating that "non-lethal" coercion can trigger international condemnation even when not treated as per se unlawful in every circumstance.[112] Likewise, *Guyana v. Suriname* held that a coercive naval expulsion of an oil rig was unlawful, underscoring that limited force at sea can still breach legal constraints.[113]

These authorities converge on a strategic point: contested maritime encounters often turn less on maneuvering performance and more on whether restraint and necessity can be demonstrated. Autonomy without evidentiary legibility invites exploitation. If an autonomous platform cannot prove what warnings it issued, what distances were involved, or how escalation was sequenced, adversaries can contest legitimacy, claim provocation, or depict lawful conduct as reckless. This creates incentives to manufacture ambiguity through close approaches and engineered "incidents."

Assurance infrastructure reduces this vulnerability by producing a credible, time-synchronized record of signals, proximity, contact behavior, and escalation sequencing. In addition to supporting legal defensibility, such records enable faster and more credible diplomatic engagement, public messaging, and operational debriefing when incidents occur.[114] Assurance records also reinforce sovereign immunity posture by demonstrating that a sovereign platform complied with collision-avoidance rules and that any coercive steps were constrained by lawful

governance. While immunity can limit jurisdictional exposure, it does not automatically prevent strategic or reputational costs; evidentiary legibility narrows the space for misinformation and accelerates credible fact presentation.

C. Alliance and Insurer Confidence Effects

Assurance infrastructure also standardizes accountability across coalition partners and reduces uncertainty for commercial risk stakeholders. Maritime autonomy will not be fielded uniformly: different navies and companies will deploy different autonomy stacks, sensors, and remote supervision designs. Without a common evidentiary framework, coalition operations face a predictable fragmentation problem—two platforms may behave similarly, but only one may be able to prove compliance under scrutiny. In contested regions where political costs are high, that uncertainty generates alliance friction and operational hesitancy.

The MASS regulatory ecosystem points toward an emerging norm: autonomy must be governed through evidence. The IMO's regulatory scoping exercise emphasizes evaluation of MASS operations against existing international instruments, pushing states toward compliance frameworks that can demonstrate equivalence across different autonomy models.[115] National guidance further underscores the need for compliance demonstration in mixed-traffic environments, particularly for smaller autonomous vessels.[116]

Assurance infrastructure enables coalition interoperability by making evidence portable. Rather than requiring partners to adopt identical autonomy technology, an assurance operating layer can produce a shared accountability output: a synchronized record of signaling, maneuvering, and escalation sequencing that can be reviewed under common legal standards. This aligns with the reality that the COLREGs function as the universal language of maritime interaction. If autonomous platforms can demonstrate compliance with COLREG signaling and collision-avoidance obligations, coalition partners can integrate operations with reduced legal and political risk.[117]

Insurer confidence follows the same logic. Maritime underwriting and claims adjudication depend on reconstruction and attribution. Under doctrines such as *The Pennsylvania* rule, failure

to rebut a navigational breach can be dispositive.[118] Courts have also indicated that reasonable seamanship may require adoption of available safety technologies, suggesting that evidence-oriented assurance mechanisms may become expected rather than optional as risk markets adapt.[119] As autonomy expands, insurers will demand higher-quality evidence to price risk, allocate fault, and resolve claims. Assurance infrastructure supplies the evidentiary backbone necessary for that market to function.

In sum, assurance operating layers reduce escalation risk by making restraint provable, stabilize gray-zone encounters by narrowing narrative ambiguity, and enable coalition and commercial adoption by producing standardized accountability outputs.[120] The next section translates these implications into policy and procurement recommendations: if autonomy requires legibility, assurance infrastructure should be treated as a condition of deployment rather than a discretionary add-on.

VII. Policy and Procurement Implications

If maritime autonomy is to scale beyond controlled demonstrations, it must be governed through mechanisms that make safety, restraint, and legal compliance demonstrable. Autonomous platforms face a credibility constraint: navigation performance is insufficient without evidentiary legibility, and escalation governance cannot depend on human memory or after-action narrative when decision-making is distributed across sensors and software systems.[121]

The policy and acquisition implication is straightforward: assurance systems should be treated as baseline procurement requirements for autonomy-enabled maritime platforms, not discretionary add-ons. This is not a call for new regulation. It is a call to operationalize existing DoD policy and international maritime obligations through acquisition-driven design requirements.

A. Aligning Acquisition with DoDD 3000.09

DoD Directive 3000.09 requires autonomous weapons systems to be used with "appropriate levels of human judgment," reflecting the principle that autonomy does not eliminate

accountability.[122] Although maritime autonomy efforts often emphasize navigation and situational awareness rather than kinetic engagement, the directive's governance logic remains directly applicable: autonomy-enabled platforms must be fielded in ways that preserve meaningful human responsibility and enable review of consequential decisions. In practice, this creates a procurement requirement for reviewability infrastructure, particularly where platforms can generate escalation effects through warnings, deterrence actions, or other coercive behaviors below lethal force thresholds.[123]

Acquisition policy supports embedding these governance expectations as requirements. DoDD 5000.01 provides the framework for lifecycle management and authorized incorporation of safety and operational risk controls into system design and evaluation.[124] Within that structure, assurance infrastructure should be treated not as a research add-on but as a compliance enabler: it provides the technical means to demonstrate restraint boundaries, signaling conformity, and auditable human-on-the-loop decision points consistent with autonomy governance.[125]

Accordingly, procurement should operationalize DoDD 3000.09 by requiring evidence—not assumptions—that human judgment remained available and meaningful at the relevant moments. The standard should be whether the system preserves records showing when intervention was possible, what options were presented, and whether approval was exercised, withheld, or deferred. This shifts oversight from post-incident narrative to repeatable compliance review during certification, training validation, and operational deployment.

B. Assurance as a Condition of Autonomy Deployment

Assurance infrastructure should be treated as a condition of autonomy deployment because it is the practical mechanism by which compliance can be proven across three domains: international navigation law, operational escalation governance, and post-incident fault allocation regimes. First, COLREG compliance depends on signaling clarity and context-appropriate precautions. Rule 36 permits supplemental light or sound warnings when necessary, but only if they cannot be mistaken for signals authorized elsewhere in the rules.[126] Rule 2 further requires vessels to take "all necessary precautions" to avoid immediate danger, including beyond explicit rule text.[127] For

autonomy-enabled vessels, these obligations generate an evidentiary requirement: operators must be able to prove what was signaled, when it was signaled, and why the signaling choice was appropriate under the circumstances.

Second, operational escalation frameworks require sequential restraint and accountable command responsibility. The Standing Rules of Engagement and Standing Rules for the Use of Force discipline coercive action through structured warning progression and reviewable authority.[128] If unmanned vessels are to operate in congested or contested waters, they must preserve records showing warning sequences, decision boundaries, and proportionality context. Third, maritime liability doctrine imposes severe consequences for evidentiary gaps. Under *The Pennsylvania* rule, a statutory navigational breach triggers a presumption of fault unless the violating vessel proves the breach could not have contributed to the incident.[129] The duty of reasonable seamanship may also require adoption of available safety technologies, suggesting that evidence-producing systems could become a baseline expectation as autonomy expands.[130] In practice, autonomy deployment without assurance infrastructure increases operational risk and legal exposure while driving procurement inefficiency through high-cost, high-uncertainty investigations after incidents.

Accordingly, assurance should be institutionalized as a deployment gate: autonomy-enabled platforms should not be fielded unless they can produce a synchronized, audit-grade record of maneuvering, signaling, and escalation sequencing sufficient for legal and policy review.[131]

C. Implications for Future USV Programs

The procurement case is especially acute for future USV programs because unmanned platforms are likely to operate where accountability costs are highest: congested sea lanes, littoral approaches, and regions where gray-zone coercion is common.[132] The Navy's MCM USV integration illustrates both promise and exposure: as unmanned systems become embedded in deployed mission packages, collisions, interference events, or contested encounters can generate immediate operational disruption and downstream legal and strategic consequences.[133] Future programs should adopt three procurement principles.

First, assurance should be treated as platform-agnostic infrastructure that enables reviewability across vendors and mission variants rather than tying compliance to a single proprietary autonomy stack.[134] Second, assurance should be embedded in certification pathways: autonomy readiness must include evidence readiness, with required outputs that support routine operational review and post-incident adjudication. Third, assurance outputs should be coalition-compatible artifacts—portable records that enable partners to evaluate conduct under shared legal standards without requiring identical autonomy systems.[135]

In sum, autonomy without assurance is operationally brittle and strategically risky. The procurement solution is not new law but requirements alignment: deployed maritime autonomy should include an assurance operating layer as a baseline condition of fielding to demonstrate lawful signaling, controlled escalation sequencing, and reviewable human judgment consistent with existing DoD policy and international obligations.[136]

VIII. Conclusion: Autonomy Requires Legibility

Maritime autonomy is no longer speculative. Defense unmanned surface vessels and commercial Maritime Autonomous Surface Ships are moving from prototypes toward operational deployment as mission packages and regulatory frameworks evolve to accommodate new operating models.[137] Yet autonomy will remain strategically brittle if it scales faster than the governance systems needed to control it. The central finding of this paper is straightforward: autonomy without accountability is destabilizing. Autonomous vessels may optimize navigation, but if they cannot demonstrate compliance with the legal and policy constraints that govern conduct at sea, they invite miscalculation, contested narratives, and legal exposure in precisely the environments where maritime stability matters most.

Maritime governance is enforced through reconstruction. The COLREGs apply to "all vessels," require standardized signaling, and permit supplemental warnings under Rule 36 only when those signals cannot be confused with authorized signals.[138] UNCLOS further requires flag states to ensure vessels conform to generally accepted international regulations, making compliance an

obligation of sovereign responsibility rather than private discretion.[139] Operationally, escalation doctrine assumes sequential restraint and controlled authority under established rules of engagement, even below kinetic thresholds.[140]

Autonomy fractures the assumptions that make these frameworks workable. Collision accountability and escalation governance presume that human operators can explain what happened and why. Autonomy instead distributes perception and decision-making across sensors, algorithms, and remote supervision channels, increasing fragmentation and evidentiary gaps at the exact moment legal review demands clarity. These gaps are outcome-determinative. Under collision doctrine, violation of a statutory navigation rule can trigger a presumption of fault that must be rebutted through evidence.[141] In contested maritime environments, adversaries exploit uncertainty as a tactic, using harassment and coercive maneuvers to manufacture ambiguity and contest legitimacy.[142] When evidence is incomplete, even lawful restraint can be reframed as provocation or irresponsibility.

Accordingly, this paper argues that assurance infrastructure is the missing layer that makes autonomy durable. A maritime assurance operating layer—non-controlling, sensor-agnostic, legally aware, and time-synchronized—would not replace navigation systems or autonomy stacks. It would produce a unified evidentiary record capable of reconstructing signaling, maneuvering, and escalation sequencing in legally intelligible terms.[143] This operationalizes functional equivalence as an auditable reality: autonomy that can demonstrate compliance with collision-avoidance and escalation constraints even when no human bridge team is present. The difference between experimental autonomy and durable capability is not performance in controlled demonstrations. It is how well the system withstands scrutiny when incidents occur—when narratives compete, liability is contested, and strategic stability depends on credible proof of restraint. Maritime autonomy will scale only if it becomes legible. Assurance infrastructure is therefore not optional; it is the enabling condition for lawful, credible, and stabilizing autonomy at sea.

[1] Naval Sea Systems Command, U.S. Navy Announces First Mine Countermeasures Mission Package Embarked on USS Canberra (Apr. 25, 2024); International Maritime Organization, Outcome of the Regulatory Scoping Exercise for the Use of Maritime Autonomous Surface Ships (MASS), IMO Doc. MSC 103/INF.3 (June 3, 2021)

[2] Naval Sea Systems Command Announcement
[3] Convention on the International Regulations for Preventing Collisions at Sea, Oct. 20, 1972, 1050 U.N.T.S. 16. Rules 1(a), 3(a), 36
[4] United Nations Convention on the Law of the Sea, Dec. 10, 1982, 1833 U.N.T.S. 397, Art. 94(5).
[5] *The Pennsylvania*, 86 U.S. (19 Wall.) 125 (1873)
[6] Chairman of the Joint Chiefs of Staff, Standing Rules of Engagement / Standing Rules for the Use of Force for U.S. Forces (CJCSI 3121.01B) (June 13, 2005)
[7] U.S. Department of Defense, DoD Directive 3000.09: Autonomy in Weapon Systems (2012)
[8] U.S. Navy, Mine Countermeasures Unmanned Surface Vehicle (MCM USV) Fact Sheet (July 2, 2025)
[9] Naval Sea Systems Command Announcement
[10] Carter Johnson, Update on the U.S. Navy's Littoral Combat Ship Mine Countermeasures Mission Package, Naval News (Apr. 1, 2025)
[11] IMO MASS Regulatory Scoping Exercise Outcome; International Maritime Organization, Autonomous Ships: Regulatory Scoping Exercise Completed (May 25, 2021)
[12] COLREGS; UNCLOS
[13] COLREGs, Rule 1(a)
[14] COLREGs, Rule 3(a)
[15] Yoshifumi Tanaka, The International Law of the Sea (Cambridge Univ. Press 2012).
[16] Christopher C. Swain, Towards Greater Certainty for Unmanned Navigation, 3 Geo. L. Tech. Rev. 119 (2018); Henrik Ringbom, Regulating Autonomous Ships: Concepts, Challenges and Precedents, 50 Ocean Dev. & Int'l L. 141 (2019)
[17] IMO MASS Regulatory Scoping Exercise; IMO MASS Regulatory Scoping Exercise Outcome
[18] *The Pennsylvania*
[19] U.S. Department of Defense, Law of War Manual (Dec. 2016, updated July 2023); CJCSI 3121.01B
[20] COLREGs, Rule 36
[21] UNCLOS, Art. 94(3)
[22] COLREGs; CJCSI 3121.01B; DoD Directive 3000.09; U.S. Department of Defense, U.S. Department of Defense, Law of War Manual (Dec. 2016, updated July 2023).
[23] DoD Law of War Manual; CJCSI 3121.01B
[24] CJCSI 3121.01B; DoD Law of War Manual
[25] Raul (Pete) Pedrozo, Demystifying China's Gray Zone Aggression: Water Cannons, Ramming, and the Use of Force, Small Wars Journal (Aug. 6, 2025).
[26] DoD Law of War Manual; CJCSI 3121.01B
[27] DoD Law of War Manual; CJCSI 3121.01B
[28] CJCSI 3121.01B
[29] UNCLOS, Art. 301; United Nations, Charter of the United Nations, June 26, 1945
[30] U.N. Charter, Art. 51; *The M/V "Saiga" (No. 2) Case* (*Saint Vincent and the Grenadines v. Guinea*), Judgment, ITLOS (July 1, 1999)
[31] *M/V Saiga (No. 2)*; Permanent Court of Arbitration, *Guyana v. Suriname*, Award (Sept. 17, 2007).
[32] U.S. Navy MCM USV Fact Sheet; Naval Sea Systems Command Announcement
[33] COLREGs, Rules 34–36
[34] UNCLOS, Art. 94(5); COLREGs, Rule 2
[35] Pedrozo; South China Sea Arbitration
[36] CJCSI 3121.01B; DoD Law of War Manual
[37] DoD Directive 3000.09
[38] DoD Directive 3000.09
[39] U.S. Department of Defense, DoD Directive 5210.56: Arming and the Use of Force (2016); United Nations Office on Drugs and Crime, Maritime Crime: A Manual for Criminal Justice Practitioners (2d ed. 2019)
[40] DoD Directive 5210.56; U.S. Department of the Navy, SECNAVINST 5500.37: Arming and the Use of Force (2020)
[41] U.S. Department of Defense, DoD Directive 5000.01: The Defense Acquisition System (2020).
[42] IMO MASS Regulatory Scoping Exercise; Defence Safety Authority, Guide to Regulation of Maritime Autonomous Systems (DSA03-DMR) (2024).
[43] DoD Directive 3000.09
[44] DoD Directive 3000.09
[45] *The Pennsylvania*

[46] *The T.J. Hooper*, 60 F.2d 737 (2d Cir. 1932).
[47] *Oceanic Steam Navigation Co. v. Aitken*, 196 U.S. 589 (1905).
[48] UNCLOS, Art. 94(5)
[49] COLREGs, Rule 2
[50] Pedrozo; South China Sea Arbitration
[51] COLREGs, Rules 1(a), 20–36
[52] Tanaka
[53] Swain; Ringbom
[54] *The Pennsylvania*; *Oceanic Steam Navigation Co. v. Aitken*
[55] COLREGs, Rule 36
[56] COLREGs, Rule 1(e); Tianyu Xu, Jing Guan & Wenda Xia, Review on "A Vessel of Special Construction or Purpose" in Relation to Collision Prevention, 4 J. Int'l Mar. Safety, Env't Aff. & Shipping 76 (2020)
[57] *Oceanic Steam Navigation Co. v. Aitken*
[58] UNCLOS, Art. 94(5)
[59] *The Pennsylvania*; Michael N. Schmitt & David S. Goddard, International Law and the Military Use of Unmanned Maritime Systems, 98 Int'l Rev. Red Cross 567 (2016)
[60] COLREGs; CJCSI 3121.01B; DoD Directive 3000.09
[61] COLREGs, Rules 5–8; Pedrozo
[62] DoD Directive 3000.09
[63] UNCLOS, Art. 94(5); *The Pennsylvania*
[64] Schmitt & Goddard
[65] COLREGs, Rules 1(e), 36; DoD Directive 3000.09
[66] Pedrozo; South China Sea Arbitration
[67] COLREGs; UNCLOS, Art. 94(5); CJCSI 3121.01B
[68] *The Pennsylvania*
[69] UNCLOS, Art. 94(5)
[70] CJCSI 3121.01B; DoD Law of War Manual
[71] COLREGs, Rules 1(a), 3(a)
[72] Swain; Ringbom
[73] COLREGs, Rules 20–31, 36
[74] COLREGs, Rule 2
[75] UNCLOS, Arts. 94(1), 94(3), 94(5)
[76] CJCSI 3121.01B
[77] UNCLOS, Art. 301; U.N. Charter, Art. 2(4)
[78] *M/V Saiga (No. 2)*; *Guyana v. Suriname*
[79] DoD Directive 3000.09
[80] IMO MASS Regulatory Scoping Exercise
[81] COLREGs
[82] *The Pennsylvania*
[83] UNCLOS, Art. 94(5); CJCSI 3121.01B
[84] Swain; Schmitt & Goddard
[85] COLREGs
[86] CJCSI 3121.01B; DoD Directive 3000.09
[87] Schmitt & Goddard
[88] IMO MASS Regulatory Scoping Exercise; Defence Safety Authority Guide to Maritime Autonomous Systems; Maritime and Coastguard Agency, MGN 702 (M) Amendment 1: Maritime Autonomous Surface Ships (MASS) of Less Than 2.5 Metres in Length Overall (Mar. 19, 2025)
[89] *The Pennsylvania*
[90] COLREGs, Rule 36
[91] COLREGs, Rule 1(e); Xu et al.
[92] CJCSI 3121.01B
[93] COLREGs, Rule 36; CJCSI 3121.01B
[94] *The Pennsylvania*
[95] Pedrozo; South China Sea Arbitration
[96] UNODC Maritime Crime Manual
[97] *M/V Saiga (No. 2)*; *Guyana v. Suriname*

[98] DoD Directive 3000.09
[99] DoD Directive 3000.09; CJCSI 3121.01B
[100] COLREGs; UNCLOS, Art. 94(5); DoD Directive 3000.09
[101] CJCSI 3121.01B; COLREGs; UNCLOS, Art. 94(5)
[102] COLREGs, Rules 2, 20–36
[103] UNCLOS, Art. 94(5)
[104] CJCSI 3121.01B
[105] *The Pennsylvania*
[106] *Oceanic Steam Navigation Co. v. Aitken*
[107] COLREGs, Rules 2, 36
[108] CJCSI 3121.01B; DoD Directive 3000.09
[109] UNCLOS, Arts. 95–96
[110] *The Schooner Exchange v. McFaddon*, 11 U.S. (7 Cranch) 116 (1812).
[111] Pedrozo
[112] South China Sea Arbitration
[113] *Guyana v. Suriname*
[114] COLREGs, Rule 36; UNCLOS, Art. 94(5)
[115] IMO MASS Regulatory Scoping Exercise Outcome; IMO MASS Regulatory Scoping Exercise
[116] MCA MGN 702
[117] COLREGs; Tanaka
[118] *The Pennsylvania*
[119] *The T.J. Hooper*
[120] CJCSI 3121.01B; COLREGs; UNCLOS, Art. 94(5)
[121] CJCSI 3121.01B; COLREGs; DoD Directive 3000.09
[122] DoD Directive 3000.09
[123] CJCSI 3121.01B; DoD Directive 5210.56
[124] DoD Directive 5000.01
[125] DoD Directive 3000.09; COLREGs
[126] COLREGs, Rule 36
[127] COLREGs, Rule 2
[128] CJCSI 3121.01B
[129] *The Pennsylvania*
[130] *The T.J. Hooper*
[131] UNCLOS, Arts. 94(3), 94(5); COLREGs; CJCSI 3121.01B
[132] Pedrozo
[133] Naval Sea Systems Command Announcement; U.S. Navy MCM USV Fact Sheet
[134] IMO MASS Regulatory Scoping Exercise
[135] Tanaka; MCA MGN 702
[136] DoD Directive 3000.09; COLREGs; UNCLOS, Art. 94(5)
[137] Naval Sea Systems Command Announcement; IMO MASS Regulatory Scoping Exercise
[138] COLREGs, Rules 1(a), 3(a), 36
[139] UNCLOS, Art. 94(5)
[140] CJCSI 3121.01B
[141] *The Pennsylvania*
[142] Pedrozo; South China Sea Arbitration
[143] DoD Directive 3000.09; COLREGs