

AI Operational Guidelines for National Security: A Proposed Field Guide for Capability-Specific Governance

Vanderbilt University Institute of National Security Wicked Problems Lab

Contact

Author

Ian Miller

Lab director

Brett Goldstein

Project inquiries and data contribution wpl@vanderbilt.edu

Executive Summary

The United States faces a mismatch between its strategic vision for artificial intelligence (AI) in national security and the operational reality of deploying AI systems. The National Security Commission on Artificial Intelligence articulated a comprehensive strategic imperative for American technological leadership and rapid AI adoption across defense and intelligence agencies. The Department of Defense formalized this with five ethical principles for AI development [5]: systems must be responsible, equitable, traceable, reliable, and governable. Yet for military planners, acquisition professionals, and operational commanders, these frameworks provide limited guidance on how to actually implement AI systems in practice.

This gap manifests in several ways. A program manager tasked with acquiring an AI-enabled intelligence analysis tool needs to know: What testing is required before deployment? How should humans interact with the system in operational settings? When the underlying model is updated, what revalidation is necessary? What human decisions cannot or should not be delegated to automation? Existing policy documents do not provide answers to these questions.

The problem extends beyond implementation difficulty. Current policy treats artificial intelligence as a single technology. In practice, the systems deployed across military, intelligence, and cyber operations exhibit fundamentally different capabilities, risks, and governance requirements. A narrow machine learning classifier trained to detect objects in imagery operates under entirely different constraints and enables different workflows than a large language model used to summarize intelligence assessments. An autonomous system that recommends resource allocation differs materially from an agentic system that executes cyber defense operations within defined boundaries. Applying identical governance frameworks to these disparate capabilities either constrains low-risk systems unnecessarily or fails to adequately address high-risk ones.

Existing frameworks also share a deeper weakness: they are largely assertion-based rather than evidence-based. They state that AI systems should be tested, that humans should exercise meaningful judgment, and that performance should be monitored for drift, but they do not report what testing actually reveals, at what thresholds human oversight measurably improves outcomes, or how quickly deployed model performance degrades under realistic operational conditions. Practitioners are asked to implement requirements whose costs, benefits, and failure rates have not been quantified for the systems, tasks, and conditions in which they will actually be used. Relevant measurement exists in adjacent research literatures, but it has not

been collected, contextualized, or connected to the decisions a program office has to make. The result is guidance that cannot be argued with, refined, or falsified, and that consequently carries little weight with the engineers and program managers who must act on it.

Additionally, existing frameworks face an inherent structural limitation: they are static documents in a rapidly evolving technical landscape. The Department of Defense’s guidance on autonomous weapons, issued in 2023, had its previous substantial update in 2012. The intervening period saw the emergence of large language models, multimodal systems, reasoning models, and agentic AI frameworks that pose governance questions entirely unanticipated by that policy. For another example, the National Security Commission on Artificial Intelligence’s final report, while comprehensive, was published before ChatGPT’s public release and thus predates the very frontier capabilities now driving national security applications of AI. Policy updates, when they occur, move on bureaucratic timelines measured in years. Technical capabilities advance on timelines measured in months or even weeks.

This mismatched pace creates downstream consequences. Acquisition programs face uncertainty about acceptable testing and validation practices, leading to delayed fielding of systems that could enhance operational effectiveness. Operators deploy systems with inadequate human-AI teaming guidance, creating risks of both over-automation and failure to fully utilize available capabilities. The absence of implementation-grade standards means that different programs apply different governance models, reducing interoperability and preventing institutional learning from operational experience.

Finally, existing frameworks inadequately address the competitive context in which U.S. national security AI deployment occurs. While American policy emphasizes governance constraints alongside capability development, strategic competitors approach AI integration with far fewer constraints. China’s military doctrine explicitly incorporates AI-enabled operations as central to future warfare, without comparable ethical frameworks limiting deployment. Russia has deployed AI-enabled targeting and information operations in Ukraine, albeit with limited success [4]. Framing responsible AI governance as purely a matter of constraint misses the strategic reality: clear operational guidance can enable faster, more predictable, more interoperable systems, which are precisely the attributes needed to maintain technological advantage.

We propose a framework to address these gaps. Rather than restating high-level principles already established in existing policy, this framework translates those principles into capability-specific operational guidance grounded in measured performance data. It does so through four mechanisms.

First, we develop a taxonomy of AI capabilities actually deployed in national security contexts: classification and detection systems, autonomous recommendation systems, foundation models applied to intelligence analysis, agentic systems operating within defined boundaries, and generative systems producing novel content. For each category, we specify the governance model appropriate to its risk profile, the testing and validation practices required before operational deployment, the human-machine teaming structure that enables both effectiveness and control, the mechanisms for post-deployment updates and revalidation, and the failure modes and safeguards required.

Second, and central to this project’s contribution, we ground each capability module in original empirical work rather than assertion. Vanderbilt Institute of National Security Wicked Problems Lab (WPL) and our collaborators will generate, contribute, and publish quantitative evidence alongside the guidance itself. This includes:

- **Model evaluation** on tasks relevant to national security, reporting accuracy and unsupported-assertion rates rather than vendor-supplied claims.

- **Adversarial and red-team results** characterizing robustness to prompt injection, data poisoning, and evasion, and the residual risk that remains after mitigation.
- **Human-machine teaming experiments** measuring automation bias, error-catch rates, time-to-decision, and analyst workload across oversight configurations, so that “human-in-the-loop” becomes a quantified design choice rather than a slogan.
- **Drift and revalidation data** documenting how quickly fielded performance degrades and what monitoring cadence detects it in time.
- **Acquisition and program data** on testing timelines, revalidation costs, and fielding delays attributable to ambiguous guidance.

Third, we establish a mechanism for regular, ongoing updates to reflect frontier capability development. Rather than waiting for the next strategic review or policy cycle, we propose a three-month update cadence tied to documented advances in AI capabilities, newly collected evaluation and operational data, lessons from field experience, and evolving threat assessments. Each update includes versioning, detailed changelogs, archived prior versions, and a record of which recommendations changed because new evidence contradicted earlier findings, allowing practitioners to understand what has changed and why.

Fourth, we pose this project as a collaborative effort involving military acquisition professionals, operational commanders, defense technologists, AI researchers, and policy experts. Collaborators contribute not only judgment, but also data: evaluation results, test findings, and operational measurements that no single institution could assemble alone. This ensures our guidance reflects both technical reality and operational necessity, and remains durable across political transitions by grounding itself in capability, measurement, and mission effectiveness.

The timing for such a framework is urgent. The frontier AI capabilities now available, particularly large language models and multimodal systems, create operational possibilities and governance questions that existing frameworks do not address, and for which almost no publicly available performance data exists in a national security context. The current administration has signaled commitment to competitive speed in AI development and deployment. This framework demonstrates that responsible governance and competitive pace are not contradictory imperatives. Clear, empirically grounded operational standards, far from slowing deployment, can accelerate acquisition and fielding by reducing uncertainty and preventing the false starts that come from inadequate initial guidance.

The remainder of this document outlines the problem in detail, presents the proposed framework and its supporting evidence base, and describes the mechanism for its ongoing development and maintenance.

1 The Gap: Why Existing Frameworks Fall Short

The United States does not lack AI policy documents. Since 2019, the Department of Defense, the National Institute of Standards and Technology, the State Department, and two administrations have all issued guidance on the development and use of artificial intelligence for national security purposes. The problem is not absence but specificity: nearly all of this material operates at a level of abstraction that cannot be executed by the people responsible for buying, testing, fielding, and operating AI systems. Where documents do descend to specifics, they typically address a narrow slice of the problem, such as autonomous weapons, a single mission area such as public affairs, or generic enterprise risk management, and leave the capabilities that dominate current national security use, particularly foundation models and agentic systems, largely unaddressed.

This section maps that landscape and identifies six distinct failure modes. Some are failures of specificity, some of structure, and one an absence of published evidence, which is a failure that undermines the credibility of everything built on top of it.

1.1 The Existing Landscape

Table 1 summarizes the principal documents that constitute current U.S. and allied guidance on AI in national security.

Taken together, these documents establish three things well. They articulate a strategic case for urgency. They establish norms that are broadly accepted in the US and among allies. And, in cases of the NIST and MITRE materials, they provide useful vocabulary for describing risks and attacks. None of that work needs to be redone, and this field guide does not attempt to redo it.

What the landscape does not provide is a path from those commitments to a fielded system. Six gaps account for the shortfall.

1.2 The Implementation Gap

The five DoD AI ethical principles are sound commitments, but unusable as specifications. “Reliable” does not state a performance floor, a test set, or a confidence interval. “Traceable” does not specify what must be logged, at what granularity, for how long, or in what format. “Governable” does not indicate whether a system requires a hardware interlock, a software kill switch, a human approval gate, or merely a documented shutdown procedure. Each term admits a range of implementations spanning orders of magnitude in cost and risk, and existing guidance offers no basis for choosing among them.

Consider the actual decision sequence facing a program manager acquiring an AI-enabled intelligence analysis capability. They must specify an acceptance threshold for model performance. They must define what constitutes a material model change requiring revalidation, as opposed to a routine update they can accept without retesting. They must decide whether analysts will see model confidence scores, and if so, whether those scores are calibrated well enough that displaying them improves rather than degrades analyst judgment. They must determine what the system does when it encounters inputs outside its training distribution, and whether that behavior is detectable by the operator. They must establish who signs off on deployment, and based on what evidence.

The existing corpus answers none of these. The NSCAI report tells them that AI should be governable. Directive 3000.09 addresses weapon systems and does not reach analysis tools. The NIST AI RMF gives them

Table 1: Principal existing frameworks on AI in national security

Document	Source	Date	Core focus
AI Ethical Principles [5]	DIB → DoD	2020	Five principles: responsible, equitable, traceable, reliable, governable
NSCAI Final Report [14]	NSCAI	2021	Comprehensive strategy for AI competitiveness
Responsible AI Strategy & Implementation Pathway	CDAO	2022	Translates the five principles into lines of effort
Directive 3000.09 [6]	DoD	2023	Autonomy in weapon systems, human judgment requirements
AI Risk Management Framework[10]	NIST	2023	General-purpose risk process: govern, map, measure, manage
Political Declaration on Responsible Military Use of AI	Dept. of State	2023	Voluntary international norms, multi-state endorsement
EO 14110[2]	White House	2023	Broad AI governance, dual-use model reporting, red-teaming
Data, Analytics, and AI Adoption Strategy [7]	DoD	2023	Adoption pathways and enterprise data posture
AI 600-1 (Generative AI Profile) [11]	NIST	2024	Generative-AI-specific risks mapped to the AI RMF
AI 100-2 (Adversarial ML) [12]	NIST	2025	Taxonomy of attacks on AI systems
Executive orders on AI [3]	White House	2025	Revocation of EO 14110, deregulation and competitiveness emphasis
TEVV guidance for AI and autonomy	CDAO / USD (R&E)	ongoing	Test and evaluation approaches for AI-enabled systems
ATLAS [1]	MITRE	ongoing	Adversarial threat landscape for AI systems
DoDM 5000.101 [8]	DOT&E	2024	Test evaluation of AI-enabled systems
DoDI 5400.19 [9]	ATSD(PA)	2025	AI in public affairs: generative AI notice, labeling, and content provenance requirements
AI Cybersecurity Risk Management Tailoring Guide [13]	DoD CIO	2025	Maps ATLAS threat vectors to CNSSI 1253 control priorities across the AI system lifecycle

a process (map, measure, manage, govern), without telling them what to measure or what value constitutes success. The CDAO Responsible AI toolkit moves closer to practice but remains organized around principle compliance rather than capability-specific engineering requirements. DoD Instruction 5400.19 illustrates the pattern precisely. Before a generative capability may be used in public affairs work, the instruction directs that the tool be tested and monitored to validate that it meets acceptable levels of performance against identified measures. It does not state what those levels are, which measures identify them, or what evidence satisfies the requirement. The direction is sound and the practitioner is still left to supply the thresholds it presumes. Faced with this, program offices default to one of two failure modes: they either impose maximal testing burdens that delay fielding of low-risk capability, or they accept vendor assurances that no independent party has evaluated.

Forcing each program manager to manage this hurdle individually is a recipe for turning well-intentioned principles into messy and inconsistent operational outcomes.

1.3 The “AI-as-a-Monolith” Problem

Existing frameworks treat “AI” as a single technology class. The systems actually deployed across military, intelligence, and cyber operations differ so substantially in failure mode, oversight requirement, and testing need that common governance is incoherent. Table 2 illustrates the divergence.

Table 2: Governance requirements across capability categories

Capability	Example	Primary failure mode(s)	Oversight model	Testing need
Classification / detection	Object detection in ISR feeds	Accuracy failure	Analyst reviews outputs	Standard ML validation & monitoring
Autonomous recommendation	Target or resource prioritization	Systematic ranking bias leading to misallocation	Human selects from ranked options	Stress testing & adversary scenario play
Foundation model for analysis	LLM summarizing reporting	Hallucinations, Unsupported inference, & Overreliance	Analyst verifies against sources	Adversarial testing & citation checks
Agentic system	Autonomous cyber defense	Uncontrolled action chain causing escalation & poisoning leading to data breach	Bounded ability, human approval at defined gates	Extensive sandboxing & interlocks
Generative / influence	Synthetic media creation	Misinformation, Leaks, Overreliance & Provenance Degradation	Notice and labeling, durable content credentials, authorization and release policy	Red-teaming, usage impact study & adversary capability modeling

The practical consequence is misallocation of governance effort in both directions. A classifier that flags candidate objects for analyst review, and whose errors are caught within seconds by the reviewing analyst, does not require the oversight apparatus appropriate to a system that takes network actions autonomously. Applying the latter’s requirements to the former delays capability for no risk reduction. Conversely, an agentic system operating at machine speed inside a live network cannot be governed adequately by a “review-the-output” plan that suffices for a classifier, because by the time a human reviews the output the action has already occurred. A single framework calibrated to the midpoint of these extremes fails in both respects.

Directive 3000.09 is the clearest illustration. It provides specific guidance on human judgment in the use

of force, and it is the most operationally useful document in the corpus for the narrow class of systems it addresses. It was not written to govern foundation models used for analytic production, agentic systems used for network defense, or generative systems used to produce content, and it does not do so. One narrow exception exists, and it is instructive. DoD Instruction 5400.19 governs the use of artificial intelligence in public affairs activities, and it is highly specific: it sets notice requirements for generative outputs, directs that photo-realistic depictions be labeled as photo illustrations, requires that the extent of generative use be recorded in product metadata, and supplies decision matrices a practitioner can actually follow. Its applicability, however, is bounded by mission area rather than by capability, and it states expressly that it does not apply to the use of artificial intelligence capabilities for non-public-affairs activities. Generative use in intelligence production, planning support, and training therefore remains outside it, as do foundation models and agentic systems in every mission area. Those capabilities constitute the bulk of near-term national security AI adoption.

DoD Manual 5000.101, which governs operational test and evaluation for AI-enabled and autonomous systems, likewise excludes generative AI and reinforcement learning from its scope by its own terms, so the vacuum extends to test policy as well as to use-of-force policy.

1.4 The Evidence Deficit

The preceding gaps are widely acknowledged. A third is not, and it is more damaging. The existing corpus is assertion-based rather than evidence-based.

Existing frameworks rightly state that AI systems should be tested before deployment. They do not report what testing reveals: what accuracy foundation models achieve on national-security-relevant tasks, how frequently they produce unsupported assertions, whether their expressed confidence corresponds to actual correctness, or how these figures vary across model generations and prompting strategies.

They rightly state that humans should exercise meaningful judgment over AI outputs. They do not report at what point human review measurably improves outcomes, how often reviewers actually catch model errors under realistic workload and time pressure, how quickly automation bias sets in as apparent system reliability increases, or what interface and workflow designs preserve genuine scrutiny rather than nominal sign-off. “Human-in-the-loop” functions in the current literature as a reassurance rather than a specification, and its efficacy is assumed rather than measured. The Department’s own Responsible AI Strategy and Implementation Pathway concedes as much, stating that there is currently no method to measure user trust in an AI system or to measure what factors influence a user’s trust decisions.

They rightly state that deployed systems should be monitored for performance drift. They do not report observed drift rates, the operational conditions that accelerate degradation, or the monitoring cadence required to detect meaningful degradation before it affects mission outcomes. DoD’s own cybersecurity tailoring guidance concedes the point, noting that standards for monitoring frequency and pre-established model efficacy thresholds will likely still need to be established by the Chief Digital and Artificial Intelligence Officer and individual system owners.

They rightly state that AI systems face adversarial threats. NIST AI 100-2 and MITRE ATLAS usefully enumerate attack classes, and the DoD Artificial Intelligence Cybersecurity Risk Management Tailoring Guide goes further than either, mapping threat vectors drawn from ATLAS to specific control families that organizations should apply at each phase of the AI system lifecycle[13]. None of the three reports success rates against realistic defended configurations, nor the residual risk remaining after recommended mitigations are applied, which is the information a program manager actually needs in order to make a fielding decision.

The tailoring guide is candid about this, describing its own mapping as a “reasonable starting point” rather than a complete analysis of the threat area, and stating that it “does not establish AI system performance expectations”.

The consequence is that the guidance cannot be argued with, refined, or falsified. A recommendation without evidence cannot be tested against experience, cannot be revised when conditions change, and cannot be defended against a program office arguing that the testing is unnecessary. It also cannot be properly prioritized, meaning that, absent measurement of which safeguards catch which failures at what cost, every safeguard appears equally warranted. In practice, this usually means that the cheapest ones are implemented and the rest are ignored.

This deficit has an understandable cause. Insofar as this evaluation data exists, much of it is classified, proprietary, or held in program-specific test records that were never intended for publication. The result is that the open literature on which practitioners and policymakers rely contains almost no quantitative characterization of how AI systems perform on national security tasks under national security conditions. That is the gap this project treats as its primary contribution, and Section 2 describes the measurement program intended to close it.

1.5 Static Documents in a Dynamic Field

Every document in Table 1 is a point-in-time artifact, and most lack any defined mechanism for revision.

The timeline for AI technology is unforgiving. Directive 3000.09 was originally issued in 2012 and reissued in 2023. Between those dates, the transformer architecture was introduced, scaled, and productized; large language models moved from research curiosity to routine deployment; multimodal and reasoning models emerged; and agentic frameworks capable of multi-step autonomous action became commercially available. The NSCAI final report, the most comprehensive strategic treatment in the corpus, was published before the public release of ChatGPT and more than a year before the release of Claude and Gemini. Its analysis of AI in national security therefore cannot incorporate what has been learned from the mass adoption of generative systems: not only their technical capabilities, but their behavioral effects on the humans who use them.

That last point deserves emphasis, because it is the kind of finding that only operational experience produces. Using a language model to draft routine correspondence carries consequences that don’t reach far beyond the tone of an email. Using the same class of tool to generate analytic alternatives, draft assessments, and inform decisions introduces a different category of risk: the gradual substitution of model output for actual analytic reasoning, in ways that are difficult to detect precisely because the output is confident and plausible. This risk was not well characterized when current policy was written, nor is it fully characterized now. It will be characterized only through sustained measurement of how analysts and operators actually behave when working alongside these systems, which is one of the empirical commitments described below.

A larger structural problem compounds the problem of substance. Policy revision moves on bureaucratic timelines measured in years, gated by coordination requirements, legal review, leadership attention, and political division. Frontier capability advances on timelines measured in months. Any document without a defined update cadence is therefore not only likely to become outdated, but guaranteed to, as its authors have no mechanism to correct the drift.

Political durability aggravates this further. EO 14110 was revoked immediately after a change in administration[3], and its substantive content, including provisions on dual-use model evaluation and red-teaming with direct national security relevance, was discarded along with its political branding. Guidance

tied to a specific administration inherits that administration's lifespan regardless of technical merit. The NSCAI report survives partially because it issued from a commission rather than executive action, but it survives as a historical document rather than a living one.

1.6 Acquisition and Implementation

Strategy documents are written for the policy community. Capability reaches the field through the acquisition system, and the existing corpus barely engages it.

A program manager operating under defense acquisition authorities needs specifications, not just principles. Specifically, they need to know how to write a statement of work for a system whose behavior changes after delivery; how to structure acceptance criteria for a capability whose performance is probabilistic rather than deterministic; how to evaluate vendor performance claims, which are typically reported on public benchmarks bearing little resemblance to the mission task; how to obtain sufficient rights to model artifacts, training data provenance, and evaluation results to permit independent government testing; how to handle model updates in a classified environment where the vendor cannot see the operational data; and how test, evaluation, verification, and validation practices designed for deterministic systems apply to systems that are retrained on new data.

Existing guidance addresses almost none of this. TEVV is close, but it is not yet expressed as capability-specific acquisition-ready requirements. DoD Manual 5000.101 directs that AI and machine learning models, human interaction with AI and autonomous systems, and learning systems that change over time be evaluated[8]. That direction governs how testing is planned and conducted, however, and does not supply the quantitative thresholds or evidence standards a program office needs at the moment it writes acceptance criteria into a requirements document. Its applicability is also narrower than its title suggests. The manual focuses on supervised learning as the most mature area for operational test and evaluation, and states expressly that it does not apply to reinforcement learning, generative AI, or other advanced types of AI, and that it will be updated when test and evaluation procedures for those applications mature. Three of the five capability categories set out in Section 2.3 — foundation models applied to analysis, agentic systems operating within defined boundaries, and generative and influence systems — therefore fall outside its scope by its own terms.

The practical result is that each program office improvises. Where issued policy governs how testing is planned and conducted but does not supply the thresholds and evidence standards a requirements document must contain, and where three of the five capability categories fall outside the scope of that policy altogether, the individual program office is left to invent its own acceptance criteria. Improvisation produces inconsistency across programs, and inconsistency prevents institutional learning. Because no two programs specify or measure the same way, no program's results can inform any other's, and the enterprise absorbs the same evaluation cost repeatedly without accumulating transferable knowledge.

Interoperability suffers as well. Systems validated against incompatible standards cannot be composed confidently, and allied systems validated against no shared standard at all cannot be integrated into combined operations without bespoke assessment each time.

1.7 Adversary Asymmetry

Existing frameworks are oriented almost entirely toward governing U.S. use. They pay little attention to four questions that directly affect operational decisions.

First, how are adversaries actually deploying AI? Chinese military writing on modern warfare treats AI-enabled operations as central to future conflict, and does so without an equivalent public framework of ethical constraint. Russia has employed AI-enabled targeting support and information operations in Ukraine, with results that appear operationally limited but that generate learning and iteration. Both state and non-state actors are experimenting with generative systems for influence and automated tooling for cyber operations. While awareness of these threats is not lacking, information on how they impact US operations is limited.

Second, how can the US counter adversary AI systems? Guidance on protecting U.S. systems from adversarial manipulation exists in taxonomy form. Guidance on identifying, characterizing, and degrading adversary AI-enabled capability is largely absent from the unclassified corpus.

Third, how can the US handle the disparity between constrained and unconstrained implementation? The relevant question is not whether the United States should abandon constraints. It should not, as the constraints have operational value. The question is which constraints beget meaningful risk reduction and which impose cost without benefit. Answering that requires measuring what each safeguard actually catches. Without measurement, the debate reduces to competing assertions about whether responsibility helps or hurts.

Fourth, what is the risk of not deploying? Every framework in the corpus treats AI deployment as the risky action and non-deployment as the safe default. In a competitive environment, however, this is incomplete. Declining to field a capability that could potentially improve decision speed, detection rates, or defensive coverage carries its own risk, which no existing framework attempts to characterize.

Framing responsible governance as pure constraint therefore misses the strategic point. Clear, specific, evidence-based operational guidance produces systems that are faster to field, more predictable in behavior, and more readily interoperable with allied capability, which are precisely the attributes required to sustain technological advantage.

1.8 Summary

Table 3 maps each identified need to its current source and the residual gap this field guide addresses.

The following four residual gaps define the scope of the proposed field guide: capability-specific requirements, published performance evidence, a defined update cadence, and technical-policy translation. The remainder of this document details how our proposed framework would address them.

Table 3: Gaps in existing literature and what the proposed field guide contributes

Need	Currently Met by	Remaining Gap
High-level ethical principles	DoD AI principles, NATO, State Dept. declaration	Adequately covered, not duplicated here
Comprehensive strategic vision	NSCAI final report	Predates frontier capability, no update mechanism
Generic risk taxonomy	NIST AI RMF, AI 600-1, MITRE ATLAS, DoD AI Cybersecurity RM Tailoring Guide	Threats mapped to controls, but no measured attack success rates or residual risk after mitigation
Autonomous weapons policy	Directive 3000.09	Limited in scope
Capability-specific requirements	Partial (DoDM 5000.101, supervised learning and autonomy; DoDI 5400.19, public affairs only)	Core contribution; no coverage of generative, agentic, or reinforcement-learning systems outside public affairs
Published performance evidence	Largely absent	Core contribution
Adversary AI assessment	Largely classified	Unclassified synthesis with operational implications
Acquisition and contracting integration	Partial (SCSP, USD(R&E) TEVV)	No published evidence on testing cost, timeline, or revalidation burden
Defined update cadence	Nothing in the corpus	Core contribution
Technical-policy translation	Limited	Core contribution

2 The Proposed Framework

The gaps identified above are addressable, but not by another simple statement of principles. What practitioners lack is a translation layer: a document that takes the commitments already established in U.S. policy, sorts AI systems into categories that actually differ in how they fail, and states for each category what to test, what to measure, how to structure human oversight, and what the measured consequences of those choices are. This section describes that framework, including its design commitments, its capability taxonomy, the guidance structure applied to each capability, and the data collection effort that grounds it. Section 3 describes how the field guide would be written and maintained.

2.1 Design Commitments

Four commitments define the field guide and distinguish it from the existing corpus. Each is stated so that a reader can determine whether we have met it.

Specific: Guidance is organized around five capability categories that differ in autonomy of action, breadth of task, observability of failure, and speed relative to human review. No requirement is stated at the level of “AI systems” generally.

Verifiable: Every substantive recommendation carries an evidence tag indicating whether it rests on measurement we or our collaborators have published, on inference from adjacent measured findings, or on judgment pending measurement. Recommendations in the third category are marked as open questions rather than presented as settled practice.

Dynamic: The field guide is published on a three-month cycle with changelogs, archived prior versions, and an explicit record of recommendations that may have changed due to new evidence contradicting earlier findings.

Translational: The document is written to be technically accurate enough that an engineer recognizes it as correct and operationally framed enough that a staff officer or program office finds it actionable, with a glossary bridging both vocabularies.

2.2 Core Principles

This field guide adopts the five DoD AI ethical principles as given. No need exists for yet another principles framework. Our contribution sits one level below: specifying what evidence would demonstrate that a given system satisfies each principle in a given capability category. Table 4 states the translations in general terms, and the individual capability modules make these more specific.

Table 4: Translation of Principle to Measurable

Principle	Stated Commitment	What Would Demonstrate it
Responsible	Humans exercise appropriate judgment and remain accountable	A named deployment authority, a stated evidence threshold for approval, and a documented decision record
Equitable	Unintended bias is minimized	Performance reported across the operational conditions and input populations the system will encounter, with disparities quantified
Traceable	Methods, data, and design are auditable	A logging specification: which inputs, model versions, outputs, and human actions are recorded, at what granularity, with what data retention
Reliable	Safety, security, and effectiveness are assured across the lifecycle	A stated performance floor on any given task, measured under an established protocol, with a defined drift-monitoring system
Governable	Systems can be disengaged or deactivated on evidence of unintended behavior	An identified interlock mechanism, a measured time-to-halt, and a tested rollback procedure

2.3 Capability Taxonomy

We propose a high-level classification of new and existing AI models into one of five categories. This makes it clear and easy to create different guidelines for very different technologies, even if they all fall under the “AI” umbrella. This list of categories is meant to be comprehensive with respect to currently existing technologies, but that is not to say that it will never need expansion in the future.

Categories are defined by governance-relevant properties rather than by model architecture, since an architecture-based taxonomy ages with each model generation. A taxonomy keyed to whether a system acts or only advises, whether its failures are visible to the operator, and whether it operates faster than human review can keep pace, survives the evolution of the underlying technology. Each category below states its examples, defining characteristics, primary failure modes, and oversight strategy.

It should be noted that many systems can and do fall into multiple categories. Most agentic systems come attached to a foundation model, for example. Systems which fall into multiple categories should be decomposed at handoff points into their component parts until each can fit into one category. Each component is then governed by the category into which it falls, which can be determined by following the flowchart in Figure 1.

2.3.1 Classification and Detection Systems

This first class includes narrowly specified models trained on labeled data that emit a class label, score, or bounding region. Examples include, but are not limited to, object detection in images and video, signal classification, network anomaly detection, biometric matching, and synthetic content detection. The system advises but it does not take any action. A human sees every consequential output and failures are usually visible to the reviewer within seconds. The dominant failure modes are accuracy loss under distribution shift and miscalibrated confidence that misleads the reviewer about how much to trust a given output. Oversight is straightforward in theory: the analyst reviews outputs and maintains decision authority. The governance question is therefore not whether humans are involved but whether review can remain genuine and feasible as volume increases and apparent reliability rises. The general machine learning literature supplies ample validation methods for such systems, but drift rates and analyst error-catch rates at realistic throughput are not published for national security contexts. Both are planned measurement targets.

2.3.2 Autonomous Recommendation Systems

Systems that rank, prioritize, or allocate, including for resource and logistics optimization, vulnerability prioritization, and courses-of-action generation. The system does not execute, but it guides the human's choice set, and options excluded from the ranking are effectively unavailable regardless of retained veto authority. The dominant failure modes are systematic ranking bias that misallocates effort consistently rather than randomly, and brittleness under conditions absent from training. Oversight nominally consists of human selection from ranked options; whether that constitutes meaningful control depends on whether the operator can see why an option is ranked as it is and whether override remains practical under time pressure. Override rates, the frequency with which operators depart from system rankings, and the conditions under which they should are not well characterized for national security decision settings in the open literature. This will likely be an open question in version 1.0 of the field guide.

2.3.3 Foundation Models for Analysis

Large pretrained models, typically language or multimodal, applied to analysis: summarization of reporting, translation, entity and relationship extraction, literature and corpus review, drafting of assessments. The system advises, but its breadth means it will be applied to tasks it was never evaluated on. Its outputs are fluent, meaning errors may not be immediately apparent. The dominant failure modes are confabulation, unsupported inference presented with unwarranted confidence, and the slower corrosion described in Section 1.5, in which model output gradually substitutes for human analytic reasoning. Oversight requires verification against sources, not just review of the output, and the workflow must make that verification cheap enough to actually occur. Vanderbilt Institute of National Security holds original measurements of output veracity and unsupported-assertion rates across current-generation models, and these will be included in the first data release accompanying version 1.0 of the field guide.

2.3.4 Agentic Systems with Limited Autonomous Action

Systems that execute multi-step action sequences within defined boundaries: autonomous network defense, automated collection retasking, autonomous logistics execution. These systems act, and they act faster than post-hoc human review can constrain. The dominant failure modes are uncontrolled action chains that escalate beyond intended scope, and manipulation of inputs or tool interfaces that redirects the system's

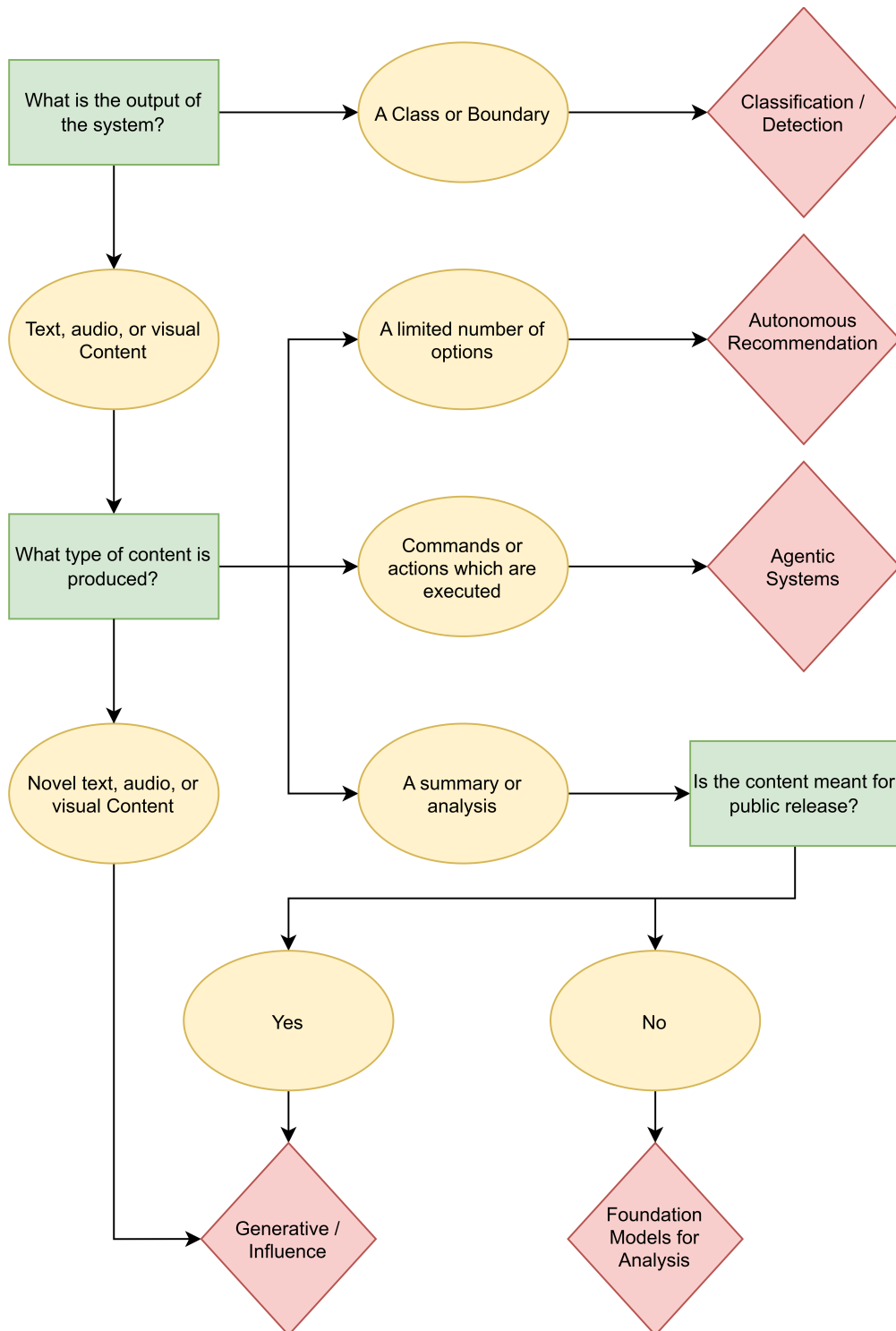


Figure 1: A simple flowchart that can be used to determine the classification of a given AI system under our framework.

authority toward an adversary's ends. Oversight cannot rest on output review due to the system's ability to effect change on its own. It requires boundaries set in advance, human approval at specified gates, comprehensive logging, and tested interlocks with measured time-to-halt. Public agentic evaluation work exists but is not contextualized for national security, and residual risk after mitigation is largely unreported. This module is the highest priority target for partnership with organizations conducting agentic capability evaluations.

2.3.5 Generative and Influence Systems

Systems whose output is novel content. This could take the form of text content, such as in automated production of public-facing materials or online influence bots. It also could take the form of synthetic image, audio, video, or other content. The category is distinguished from the others by where its risk lies. For the preceding four categories, the governing question is whether the system's output is sufficiently correct. Here, output correctness is often secondary to questions of authority and control: who authorized the content's creation, whether it can be traced to its origin, whether it remains contained to its intended use, and what happens if it does not. The dominant failure modes follow from this. Generated artifacts escape their intended context, whether through inadvertent release or intentional spread. Provenance markers degrade or are stripped, leaving no durable record of what was machine-produced. And sustained use erodes the human capability the system was meant to support, as drafting and modeling skills atrophy in ways that are invisible until the system is unavailable or wrong. As a result, oversight relies less on output review and more on policy, including explicit authorization for what may be generated and by whom, mandatory provenance marking and retention, containment requirements separating generated artifacts from collected material, and a specified release protocol for any content leaving the producing organization.

This is not hypothetical. DoD Instruction 5400.19 already imposes requirements of exactly this shape on public affairs activities, requiring notice where generative artificial intelligence is central to a product, labeling of photo-realistic depictions as photo illustrations, and recording of the purpose and extent of generative use in product metadata to safeguard digital content provenance[9]. It directs that durable content credentials be applied to visual information to the extent possible. That qualifier is where governance and evidence part: no published measurement establishes how far content credentials survive ordinary handling, re-encoding, platform ingestion, or deliberate stripping, so neither a practitioner nor a policy author can say what "to the extent possible" currently buys.

Authority questions in this category also extend beyond the technical, since content generation for influence purposes implicates legal constraints that this field guide identifies but does not attempt to resolve. The technical performance of generators is well characterized in the open literature and requires little addition from us. What is largely unmeasured is everything governance depends on: the durability of content credentials and other provenance markings under realistic handling, the rate at which synthetic material contaminates real datasets, and the severity and timescale of human skill degradation under sustained use. This data is planned for a later cycle, and will most likely be operational rather than laboratory data.

2.4 Capability-Specific Guidance Structure

Each capability module follows the same six-part structure, so that a reader who has used one module can navigate any other, and so that revisions can be made to one element without disturbing the rest. Table 5 states the elements and the questions each answers.

Table 5: Guidance structure applied to each capability module

Element	Questions it Answers
Governance model	What level of human oversight is required? Who approves deployment and updates? Which decisions cannot be delegated?
Testing and validation	Which tests are appropriate for this capability? What thresholds indicate readiness? How is performance established if an evaluator cannot fully inspect all of the data?
Integration and teaming	How are work and authority divided between human and system? What information does the operator need? Which configurations allow for proper scrutiny?
Update and maintenance	What constitutes a material change requiring revalidation? How and when is drift detected?
Failure modes and interlocks	How does the system fail? What is the fallback when it fails? How quickly can it be halted? How are failures detected and logged?
Adversary considerations	How would an adversary attack this system? Which defenses reduce risk? What residual risk remains?

Every element carries an evidence tag: **Measured**, where we or a collaborator publish supporting data; **Inferred**, where the recommendation stems from existing measurement from an adjacent context, and the data’s transference to national security is plausible but untested; or **Open**, where the recommendation reflects judgment that measurement has not yet tested. Version 1.0 will contain more **Inferred** and **Open** tags than **Measured** ones. As such, the update cycle of the project serves not only to allow evolution with the technology, but also to continually add data, correct errors, and improve confidence.

Each module is drafted in two passes. The first pass, which Version 1.0 represents, is a structured synthesis: for every element of the six-part structure, we survey the relevant published literature and state the recommendation that evidence supports, the evidence it rests on, and the limits of its transfer to a national security setting. Where the literature is thin or absent, the element is tagged **Open** and states the measurement that would close it. The second pass, distributed across subsequent cycles, replaces inference with direct measurement as our own results and partner contributions arrive. The structure does not change between passes; only the tags and the evidence behind them do. A reader can therefore see at any version which recommendations have been tested under conditions resembling their own and which have not.

2.5 Evidence Base

The evidence deficit described in Section 1.4 is what this project treats as its primary contribution. Existing frameworks assert that testing matters without reporting what testing reveals; we intend to report, at the unclassified level, with published methods and versioned data, so that others can reproduce, contest, or extend the findings. Numbers make it possible to defend and properly prioritize safeguards. The measurement program is organized around the same six questions that structure every capability module. Each element of the guidance structure has a corresponding empirical question, and the evidence tag attached to a recommendation points to the measurement that either supports it or has not yet been made. Table 6 states the correspondence and the current status of each, which doubles as the project’s research roadmap.

While we possess some data, Vanderbilt Institute of National Security seeks collaboration to fill out this table and move from majority **Open** and **Inferred** tags, to almost entirely **Measured**. This table serves as a roadmap for future experimentation by us and collaborators.

Table 6: Measurement and evidence status of each element of the guidance structure

Module element	Measurement	Evidence Status
Governance model	What evidence programs actually require before approving deployment, and what the approval process costs: testing timelines, revalidation effort, and delays attributable to ambiguous guidance	Planned
Testing and validation	Output veracity, unsupported-assertion (hallucination) rate, error distribution, and confidence calibration on analytic tasks, reported by model generation	Partially Measured
Integration and teaming	Automation bias onset, error-catch rate, time-to-decision, and workload across oversight configurations and levels of apparent system reliability	Planned
Update and maintenance	Observed performance degradation rates, conditions that accelerate drift, and the monitoring cadence required to detect degradation before it affects outcomes	Planned
Failure modes and interlocks	Measured time-to-halt, rollback success under realistic conditions, and the rate at which failures are detected by logging and monitoring rather than by consequence	Planned
Adversary considerations	Attack success rates for prompt injection, data poisoning, evasion, and other attacks on undefended and defended configurations, and residual risk after recommended mitigations	Planned

2.5.1 Methods and Reproducibility

All findings will be reported unclassified. Protocols, task definitions, scoring rubrics, and dataset versions will be published alongside results. Teaming experiments will be pre-registered where feasible. Uncertainty will be reported where possible. Null and negative results will be included on the same footing as positive ones. Evaluation artifacts will be archived with the field guide version they support, so that any recommendation can be traced to the specific measurement that justified it.

2.5.2 Contribution Model

Collaborators contribute judgment and data. Test and evaluation organizations, FFRDCs, program offices, and vendors frequently hold evaluation results that were never intended for publication but that are releasable in anonymized or aggregated form. We seek those results under terms negotiated case by case, with attribution at the contributor’s preferred level of specificity, methodology disclosure as a condition of inclusion, and contributor review before publication. Data contributors are recognized as a distinct category in the contributor list, separate from advisory reviewers.

3 Implementation

This document is a proposal. What follows describes how the field guide it proposes would be built, maintained, governed, and read, and what participation in that effort involves.

3.1 Collaborative Development

Development is organized in three tiers. Core authors set direction, decide what is published and when, and drive the update cycle, holding final editorial authority. They are named at the front of each version. Data contributors supply evaluation results, test findings, and operational measurements, and are credited within each module their data supports. Advisory reviewers comment on drafts each cycle and are acknowledged at the end of the document. Vanderbilt Institute of National Security is the field guide's permanent institutional home.

None of these roles are filled currently. This document serves as the recruitment instrument for all three, and Section 4 states how to volunteer.

Human-subjects work, including the human-machine teaming experiments described in Section 2.5, will all be conducted under Institutional Review Board review. Approved protocols will be published alongside results. This is a slower path than informal usability testing, but teaming findings are only worth reporting if the conditions under which they were obtained can be inspected and replicated.

3.2 Update Cycle and Modular Architecture

The field guide is revised on a three-month cycle. The entire document is subject to revision at each cycle, not merely the capability modules, because a taxonomy that cannot be corrected could become outdated. In practice most changes will fall in the modules, but nothing is exempt.

Revision is triggered by several inputs, including but not limited to documented advances in AI capability, newly collected evaluation or operational data, feedback and error reports from practitioners using the guidance, and reassessment of adversary use. Each cycle produces a versioned release, a changelog, and an archived copy of the prior version. Changelog entries each have a type. Of particular importance are Evidence Revision entries, recording recommendations that changed because new measurement contradicted an earlier finding. The project aims to be open and honest about errors, corrections, and new data that contradicts earlier findings.

Modules are written to be independently revisable. The canonical version is hosted by Vanderbilt Institute of National Security. All versions, changelogs, and evaluation artifacts are archived in a public repository, which also serves as a public comment channel. The specific locations of these will be announced alongside Version 1.0

3.3 Scope, Classification, and Boundaries

All findings, protocols, and guidance are unclassified and publishable. Cleared individuals may participate as authors, contributors, or reviewers, but contribute no classified information; every published claim must be independently sourceable from open material. This is a real constraint on how specific some recommendations can be, and we prefer to state it plainly rather than produce guidance whose basis a reader cannot check.

The following are outside scope. Weapons release and the use of lethal force, which Directive 3000.09 governs specifically. The field guide is a complement to that directive, rather than a competitor. Nuclear

command, control, and communications, where the governance regime is distinct and the stakes do not admit provisional guidance. Legal authorities for influence operations, which Section 2.3.5 identifies as being relevant to the generative and influence module without attempting to resolve. Classified tactics, techniques, and procedures. Contract and statement-of-work language, since producing model contractual text requires an evidence base the project does not yet have and legal expertise the team does not hold. However, this remains an open avenue for future work.

This field guide also does not direct how test organizations plan, conduct, or report operational test and evaluation; those procedures are established by issued departmental policy, and nothing here is offered as an alternative to them.

Nor does it restate or supersede the requirements that DoD Instruction 5400.19 establishes for the use of artificial intelligence in public affairs activities. Where the generative and influence module addresses the same subject matter, it is written to supply the measurement those requirements presume, such as the durability of provenance marking, the detectability of synthetic content, and the effect of notice on audience trust.

The guidance is written for U.S. national security practitioners. It is intended to be usable by allied organizations, and published protocols and shared performance vocabulary make allied systems easier to assess against a common standard, but it is not addressed to allied governments and does not attempt to account for their distinct legal and policy environments.

3.4 Audiences

The field guide serves five audiences with different needs. Table 7 indicates where each should begin.

Table 7: Audiences and recommended entry points

Audience	Primary Need	Entry Point
Program, test, and evaluation personnel	What to require and what evidence satisfies it	Relevant capability module, testing and update elements
Operational staff and commanders	How to configure oversight and what failure looks like	Relevant capability module, teaming and failure-mode elements
Policy staff	Where current frameworks leave gaps and what fills them	Section 1, taxonomy overview
Congressional and oversight staff	Whether stated safeguards are substantiated	Evidence base, evidence tags and changelog
Technical researchers	What is measured, how, and what remains open	Evidence base, protocols and open questions

Congressional and oversight readers are an important audience. Oversight of AI in national security currently depends on asserted safeguards that no published measurement substantiates. The evidence tags exist partly so that a reader outside the executive branch can distinguish a recommendation supported by data from one supported by judgment alone.

3.5 Near-Term Timeline

Version 1.0 is targeted for release in January 2027. It comprises the five capability modules, each populated across all six elements of the guidance structure as a structured synthesis of the published literature with an evidence tag and a stated measurement gap attached to every recommendation, together with one original

data release: output veracity and unsupported-assertion rates across current-generation models on analytic tasks, published with its protocol. Version 1.0 is therefore a complete instrument with one element substantiated by our own measurement and the remainder substantiated, where possible, by the field's. Evidence tags will indicate the fact that these field-substantiated modules are meant to ultimately be populated with data gathered from our lab and partners. Modules with little supporting data will be published rather than withheld, with evidence tags marking what rests on measurement and what does not, because concealing the extent of what is unknown would reproduce the deficit described in Section 1.4.

Before release, we plan to recruit core authors and advisory reviewers; complete the five modules; establish data-contribution terms with initial partners; obtain IRB approval for the first teaming protocol. At release, we will publish through Vanderbilt Institute of National Security, with the first empirical results and their protocols. Each cycle thereafter, at roughly three-month intervals, we will reassess capability and adversary developments, incorporate new data and practitioner feedback, produce a versioned release with a changelog, and archive the prior version.

A note on funding, since the measurement program raises the question. Vanderbilt Institute of National Security may fund studies conducted within Vanderbilt University. It does not offer funding to outside collaborators and is not seeking funding from them for this project. Contributions are made because the contributing organization wants the measurement to exist and to be public. This limits how fast Table 6 can be filled, and we accept that cost. A document whose findings could be attributed to a sponsor's interest would be worth less than one that fills more slowly.

4 Conclusion and Participation

The United States has articulated what responsible AI in national security should look like but has not specified how to build it. The gap is not in the intent, but in the specificity and evidence. Existing guidance operates above the level at which systems are tested, fielded, and overseen, and it asserts the value of safeguards whose costs and effects have never been measured in open literature. The consequence is that practitioners improvise, programs cannot learn from one another, and oversight rests on claims no one has publicly checked.

The field guide proposed here addresses that gap by sorting AI systems into five categories that differ in how they fail, applying a guidance structure to each, tagging every recommendation according to whether measurement supports it, publishing new measurements and their protocols, and revising the whole document every three months with a record of what changed and why. Version 1.0 will contain more open questions than settled answers. Stating where the field's knowledge ends is the first step to extending it, and the update cycle exists to close those gaps.

4.1 How to Participate

Contribute data. If your organization holds evaluation results, test findings, drift measurements, or program cost and timeline data that are releasable in anonymized form, we want them. Terms will be negotiated case by case, with your preferred level of attribution, with methodology disclosure as a condition of publication, and with your review before release.

Serve as an author or reviewer. Core authorship, module authorship, and advisory review positions are open. Module authorship suits practitioners and researchers with depth in one capability area. Advisory review is a lighter commitment of one review cycle per quarter, contributing edits and/or opinions.

Submit feedback and error reports. Once a version is published, specific disagreement is encouraged. Corrections, contrary evidence, and reports of guidance that failed in practice will be reviewed each cycle, and corrections will be recorded in the changelog.

Contact

Project inquiries & Data contribution: wpl@vanderbilt.edu

4.2 Acknowledgment

This whitepaper and all contents within would not have been possible without the generous support of the William & Flora Hewlett Foundation.

References

- [1] MITRE Corporation. Adversarial threat landscape for artificial-intelligence systems (atlas). <https://atlas.mitre.org/>. Living knowledge base, as amended; continuously updated by the MITRE Corporation. The “as amended” formulation is confirmed by the reference list of the DoD CIO Artificial Intelligence Cybersecurity Risk Management Tailoring Guide.
- [2] The White House. Safe, secure, and trustworthy development and use of artificial intelligence. Executive Order 14110, October 30, 2023; published in the Federal Register, October 2023. Executive Office of the President; revoked by Executive Order 14179 (January 23, 2025).
- [3] The White House. Removing barriers to american leadership in artificial intelligence. Executive Order 14179, January 23, 2025, January 2025. Executive Office of the President; this order revoked Executive Order 14110 (October 30, 2023).
- [4] Andrew Kramer. A drone killed three ukrainians. it was guided entirely by a.i. - the new york times, Aug 2026.
- [5] U.S. Department of Defense. Artificial intelligence ethical principles for the department of defense. Secretary of Defense Memorandum, February 21, 2020, February 2020. Issued by the Office of the Secretary of Defense; announced publicly as “DoD Adopts Ethical Principles for Artificial Intelligence,” February 24, 2020. Both dates confirmed by the reference list of DoD Manual 5000.101.
- [6] U.S. Department of Defense. Autonomy in weapon systems. DoD Directive 3000.09, January 25, 2023. Available at <https://www.esd.whs.mil/DD/>, January 2023. Office of the Under Secretary of Defense for Policy; date confirmed by the reference lists of DoD Manual 5000.101 and DoD Instruction 5400.19.
- [7] U.S. Department of Defense. Data, analytics, and artificial intelligence adoption strategy. Technical report, U.S. Department of Defense, 2023. Issued by the Chief Digital and Artificial Intelligence Office on behalf of the Department of Defense.
- [8] U.S. Department of Defense. Operational test and evaluation and live fire test and evaluation of artificial intelligence-enabled and autonomous systems. Technical Report DoD Manual 5000.101, Office of the Director, Operational Test and Evaluation, December 2024. DoD Manual 5000.101, December 9, 2024. Cleared for public release.
- [9] U.S. Department of Defense. Public affairs use of artificial intelligence. Technical Report DoD Instruction 5400.19, Office of the Assistant to the Secretary of Defense for Public Affairs, July 2025. DoD Instruction 5400.19, effective July 28, 2025. Cleared for public release.
- [10] National Institute of Standards and Technology. Artificial intelligence risk management framework (ai rmf 1.0). Technical Report NIST AI 100-1, National Institute of Standards and Technology, U.S. Department of Commerce, January 2023. Identifier and date confirmed by the reference list of the DoD CIO Artificial Intelligence Cybersecurity Risk Management Tailoring Guide.
- [11] National Institute of Standards and Technology. Artificial intelligence risk management framework: Generative artificial intelligence profile. Technical Report NIST AI 600-1, National Institute of Standards and Technology, U.S. Department of Commerce, 2024. Companion profile to the NIST AI Risk Management Framework (AI RMF 1.0).

- [12] National Institute of Standards and Technology. Adversarial machine learning: A taxonomy and terminology of attacks and mitigations. Technical Report NIST AI 100-2, National Institute of Standards and Technology, U.S. Department of Commerce, 2025. Issued by the NIST Information Technology Laboratory; earlier editions dated 2023 and 2024.
- [13] DoD Chief Information Officer. Artificial intelligence cybersecurity risk management tailoring guide. Technical report, Office of the DoD Chief Information Officer, 2025. Issued in support of DoD Instruction 8510.01, “Risk Management Framework for DoD Systems”.
- [14] National Security Commission on Artificial Intelligence. Final report. Technical report, National Security Commission on Artificial Intelligence, 2021. Congressionally chartered commission established by the John S. McCain National Defense Authorization Act for Fiscal Year 2019.