AN ANNUAL REPORT ON AI THAT HAS TO WORK INSIDE U.S. FEDERAL MISSIONS

The year the money arrived. The seat is still empty.

In fiscal 2026 the Department of Defense requested $13.4 billion for AI and autonomy, its first dedicated budget line for these capabilities, and published an AI strategy that orders frontier models deployed within 30 days of public release. Federal AI use more than doubled between 2023 and 2024. The strategy, the money, and the models have all arrived. What has not arrived, at most operational units, is the person who can carry AI across the last mile: someone who holds security engineering and AI engineering in one seat, inside the boundary, accountable for a mission outcome. This report measures that gap and describes what closing it looks like.

THE FIVE FINDINGS

— **1 · Spending is no longer the constraint.** The FY26 request allocates $13.4B to AI and autonomy. Only about $1.4B of it targets software, cross-domain integration, and AI automation, the layer where mission workflows actually change.

— **2 · Expertise is the binding constraint.** GAO found agencies struggling to locate AI specialists even to evaluate contractor bids, let alone to deliver. The talent gap now binds harder than the budget gap.

— **3 · Doctrine has converged on the forward-deployed pattern.** The January 2026 DoD AI strategy embeds AI Integration Leads in services and combatant commands and directs capability development beside operators. The department is institutionalizing what forward-deployed engineering firms have practiced for years.

— **4 · Governance is becoming a gate, not a garnish.** Auditable evidence, named human authority, and fail-closed control flows are moving from vendor slideware to acceptance criteria. Programs that cannot answer "who asked, who approved, what changed" will not clear review.

— **5 · The pilot graveyard is universal, and now it is measured.** MIT's 2025 GenAI Divide study found 95 percent of enterprise generative AI pilots produced no measurable P&L impact, and GAO reports agencies are not systematically capturing lessons from AI procurements. The same pilot dies the same death in industry and government because nothing carries the lesson between attempts.

$13.4B
FY26 DOD REQUEST, AI AND AUTONOMY

95%
ENTERPRISE GENAI PILOTS WITH NO MEASURABLE P&L IMPACT (MIT)

30 days
DIRECTED TIME FROM MODEL RELEASE TO DEPLOYMENT

88%
ORGANIZATIONS USING AI IN AT LEAST ONE FUNCTION (STANFORD)

Version 1.0, August 2026. Produced by Federal AI from public, unclassified sources, cited on the final page. Analysis and opinions are Federal AI's, offered as company positions. Free to quote with attribution. No government endorsement implied.

§1 · THE NUMBERS

Follow the money to the missing layer.

The FY26 request is a landmark: for the first time, AI and autonomy carry their own budget line. Read the composition, though, and the story sharpens. Of $13.4B requested, roughly $9.4B goes to aerial drones, $1.7B to maritime autonomy, $734M to underwater systems, and $210M to ground vehicles. Platforms dominate. The layer where a squadron's daily decisions change, software and cross-domain integration at about $1.2B plus roughly $200M for AI and automation technologies, is about a tenth of the line.

That ratio is not wrong; platforms are expensive. But it means the enterprise and mission-support layer, where most federal AI value will actually be realized this decade, competes for a thin slice. Three years earlier, DoD's direct AI support spending was about $1.1B total, and GAO was already warning that the department lacked department-wide guidance for how components should acquire AI. The money has multiplied roughly twelvefold. The guidance, the acquisition muscle, and above all the people have not kept pace.

Meanwhile demand is compounding. GAO reports federal agencies more than doubled their AI use from 2023 to 2024, and its 2026 review of AI acquisitions found a pattern that should worry every program office: agencies could not reliably find AI-literate people to evaluate bids, could not project AI costs with confidence, and were not capturing lessons from one procurement to inform the next. Four agencies, including DoD, concurred with GAO's recommendation to systematically document AI acquisition lessons. Concurrence is not capability.

WHAT THIS MEANS FOR A MISSION OWNER

If you own a recurring operational decision, the practical reading of the numbers is this: money for your problem exists, but it is not fenced for you, and the expertise to spend it well is the scarcest input in the system. The programs that move in 2026 and 2027 will be the ones that arrive with the expertise attached: a bounded workflow, a fixed price inside the simplified acquisition threshold, acceptance measures in writing, and an engineer who can sit inside the boundary. Programs that arrive as ideas will queue behind the talent gap.

§2 · THE WIDER EVIDENCE

Industry already ran this experiment. Read the results.

Two of the most cited research artifacts of the past year, MIT's GenAI Divide study and the Stanford HAI AI Index, describe the commercial economy, but their findings map onto federal missions with uncomfortable precision.

MIT: adoption is easy, transformation is rare. The MIT study found that despite an estimated $30 to 40 billion in enterprise generative AI investment, 95 percent of pilots produced no measurable profit-and-loss impact; only about 5 percent of integrated systems created significant value. The causes it names are the federal walls wearing business attire: tools that never embed into real workflows, investment aimed at visible functions rather than high-return ones, and organizations that avoid the friction of process change. Two findings deserve a mission owner's full attention. First, externally partnered, customized deployments succeeded roughly twice as often as internal builds, which is the commercial restatement of the forward-deployed thesis: the scarce ingredient is not the model but the embedded engineer who owns integration to outcome. Second, MIT documents a shadow AI economy: while only about 40 percent of companies had official AI subscriptions, roughly 90 percent of workers reported using personal AI tools for job tasks. Any leader who believes their organization has no unsanctioned AI use is measuring the sanctioned kind. GenAI.mil, the department's move to give three million personnel a governed alternative, is best read as a direct answer to exactly this finding.

Stanford: capability is broad, autonomy is thin, incidents are rising. The 2026 AI Index reports that 88 percent of organizations now use AI in at least one business function, yet agentic AI, systems that execute rather than suggest, remains in single-digit adoption across most functions. Measured productivity gains are real where the work is AI-exposed: roughly 14 percent in customer-service tasks and 26 percent in software development. Documented AI incidents rose to 362 in 2025 from 233 the year before, a 55 percent increase that lands just as agencies begin wiring AI into consequential workflows. The lesson for missions is not caution for its own sake; it is that the gap between "uses AI" and "trusts AI to act" is where all the value and all the risk now concentrate, and crossing it is a governance engineering problem, not a model problem.

The synthesis. Industry spent tens of billions learning three things the mission world can have for free: pilots without workflow integration fail at 95 percent; embedded external expertise roughly doubles the odds; and autonomy without auditable governance stalls in the single digits. A mission AI program designed against those three facts looks exactly like the pattern this report describes: one accountable seat, one bounded workflow, evidence by construction.

§3 · THE STRATEGY SHIFT

Doctrine caught up to the forward-deployed pattern.

The department's January 2026 AI strategy is the most aggressive AI policy document the Pentagon has produced. It organizes effort around seven pace-setting projects spanning warfighting, intelligence, and enterprise use, including GenAI.mil, an effort to put generative AI in front of roughly three million personnel, and Enterprise Agents, playbooks for deploying AI agents across business functions. It directs that the latest models be deployed within 30 days of public release, orders monthly progress reporting, and instructs components to submit AI hiring plans within 60 days.

Two provisions matter most for mission AI. First, the strategy explicitly directs putting AI capabilities in operators' hands and gathering feedback within days, a repudiation of the multi-year requirements cycle for this class of capability. Second, it embeds AI Integration Leads within services and combatant commands so capabilities co-evolve with warfighting concepts. Read plainly: the department has concluded that AI adoption is a person-shaped problem, and that the person must sit with the mission, not at headquarters and not at a vendor.

This is a validation of the forward-deployed thesis, and it raises the bar for everyone selling into it. When every command has an AI Integration Lead, vendors who can only ship software will be handed requirements by people who know better. The firms that thrive will be the ones whose engineers can sit on the government side of the table technically: fluent in the mission's systems, credentialed for the environment, and accountable past delivery to adoption.

THE TENSION NOBODY HAS RESOLVED

A 30-day model deployment clock and a fail-closed governance posture pull in opposite directions, and the strategy largely leaves the reconciliation to implementers. The units that will move fastest are the ones that industrialize the reconciliation: pre-approved evaluation cases, standing evidence requirements, and control boundaries designed so that swapping a model is a configuration change with a test suite, not a new authorization. Speed and governance stop being rivals exactly when the governance is engineered rather than argued.

§4 · THE LAST MILE

Pilots still die at the same three walls.

The boundary wall. Capability is demonstrated in an environment that is not the mission's environment, and the plan for crossing that gap is scheduled for later. Later is where programs go to die: the authorization conversation, the offline packaging, the identity integration, and the transfer evidence all turn out to be the actual project, discovered after the budget was spent on the demo.

The authority wall. Nobody wrote down what the AI is allowed to do, so the safe answer at every review is nothing. Programs that survive define authority as engineering from day one: read-only-first connectors, a named human approving every consequential action, an append-only evidence record, and a system that holds when authority is ambiguous. Review boards approve what they can audit.

The data wall. The governing facts of a mission live in three to five systems of record with different owners, refresh rhythms, and quality. The pilot that ingested a curated extract meets the operational reality of reconciling live sources, and the reconciliation, not the model, is the hard part. GAO's finding that agencies struggle to project AI costs is partly this wall in disguise: the model is priced, the reconciliation is not.

None of these walls is new, and that is the indictment. GAO recommended systematic lessons-learned capture precisely because the same pilot dies the same death at different agencies. The cheapest AI investment any agency can make in 2026 is writing down why the last one failed.

WHERE MISSION AI IS ALREADY PAYING

The unglamorous center of gravity is readiness and operations support: the daily reconciliation of maintenance records, schedules, and supply positions into a decision. This work is measurable (mission-capable rates, schedule effectiveness, analyst hours), it lives in systems of record rather than in classified sensor feeds, and its value arrives weekly, not at initial operational capability. In current delivery, reported outcomes for one Air Force training organization include 12 to 15 analyst hours reclaimed per week from a single reconciled readiness picture. Modest per unit, and transformative multiplied across a service's daily decisions, which is why predictive maintenance and readiness analytics keep expanding fleet-wide while flashier programs queue for authorization.

§5 · WHERE MISSION AI LANDS NEXT

Decompose the mission, then automate the steps that deserve it.

The most useful mental model in applied AI right now is task decomposition, the discipline Andrew Ng has argued for consistently: do not ask "can AI do this job," ask "which tasks inside this job can a bounded, evaluated workflow do well, and which must stay human." Missions decompose beautifully, because military work is already written down as tasks, checklists, and approval chains. Six workflows we expect to cross from pilot to operations first, all unclassified patterns:

— **Readiness reconciliation.** Maintenance records, the flying schedule, and supply positions reconciled into one defensible daily picture, with every figure traceable to its source record. The proven beachhead; reported outcomes already include double-digit analyst hours reclaimed weekly.

— **Parts and demand forecasting.** Need-by dates scored against order status and historical lead times, so the constrained part surfaces days earlier and the recovery action is priced before the schedule breaks.

— **Training and evaluation scheduling.** Syllabus progress, instructor capacity, and aircraft availability solved together, with the schedule published only after a named human approves the tradeoffs.

— **Accreditation evidence preparation.** Assembling the evidence package for a security review from the ledger the workflow already writes: who asked, who approved, what changed, drafted for the ISSO rather than by the ISSO.

— **Debrief and discrepancy capture.** Turning free-text maintenance debriefs into structured discrepancy records against the authoritative system, with the original text preserved and linked, never overwritten.

— **Logistics reconciliation.** Transportation status, warehouse records, and requisitions reconciled into arrival confidence for the items a mission is actually waiting on.

What these six share: the data lives in systems of record, the decision recurs on a clock, the value is measurable in the mission's own metrics, and none requires new autonomy. They are recommend-and-hold workflows, which is why they can clear review boards now instead of in 2028.

FOUR TRENDS SHAPING THE NEXT 18 MONTHS

— **Agentic workflows over monolithic prompts.** The gains Ng predicted from iterative, tool-using, multi-step AI workflows are showing up in evaluations and in practice; for missions this means pipelines of small, testable steps, each with its own evidence, rather than one oracle call.

— **Small and local models grow into the enclave role.** Capable open-weight models now run on hardware that fits inside a disconnected boundary, making the air-gapped deployment profile a first-class citizen rather than a compromise.

— **Evaluation-first delivery becomes table stakes.** Writing the evaluation cases before the capability, then measuring instead of demonstrating, is moving from best practice to expectation, accelerated by the 30-day model-refresh clock.

— **The authority pattern standardizes.** Read, reconcile, recommend, hold: the shape of AI that review boards approve is converging across industry and government, and vendors who cannot produce an evidence ledger will find the conversation short.

§6 · THE 2026 TOOLBOX

Tools that changed the mission calculus this year.

Engineering-level, named, and practical: these are the tools an executive can hand to a technical team for evaluation this quarter. Every item below is publicly available and verifiable; inclusion is not endorsement, and every candidate still requires the customer's own security review.

— **Open-weight models fit for the enclave.** OpenAI's gpt-oss family (Apache-2.0 licensed, 120B and 20B parameter models, the smaller of which runs on a single high-end GPU) joined Meta's Llama, Google's Gemma, and Alibaba's Qwen lines in making capable models that can live entirely inside a disconnected boundary. The practical shift: a serious enclave model is now a licensing and vetting decision, not a research project. Vet provenance, license, and evaluation results for any model regardless of who published it.

— **Serving stacks that make local models operational.** vLLM has become the de facto high-throughput serving layer for self-hosted models, and llama.cpp covers constrained and edge hardware. Together they turn open weights into an on-premises inference service a platform team can actually operate.

— **Frontier models behind the boundary.** Google Distributed Cloud's air-gapped configuration now runs Gemini fully disconnected, down to a single-server appliance form factor, and frontier models are reachable inside government-authorized cloud regions through services such as Amazon Bedrock in AWS GovCloud and the Azure Government OpenAI service. The old binary, frontier capability or disconnection, is dissolving into a deployment-profile choice.

— **A connector standard instead of connector chaos.** The Model Context Protocol (MCP) has been adopted across major AI vendors as a common way to give models governed access to tools and data. For missions this matters because one audited, read-only-first connector can now serve successive models, which is exactly what a 30-day model-refresh clock requires.

— **Evaluation harnesses worth standardizing on.** Inspect, the open-source evaluation framework from the UK AI Security Institute, and CI-style tools such as promptfoo make evaluation-first delivery practical: write the cases before the capability, run them on every model swap, keep the results as evidence.

— **The department's own on-ramp.** GenAI.mil gives DoD personnel a governed generative AI workspace, the sanctioned answer to the shadow AI economy in §2. For mission owners it is also a useful baseline: general assistance is now provided; the remaining gap, and the subject of this report, is the governed mission workflow.

PRACTICAL TAKEAWAYS, IN ORDER OF LEVERAGE

— **1 · Ask every vendor for the deployment profile, not the demo.** Cloud, on-premises, or air-gapped, with the model route named for each. A vendor without a disconnection answer is a connected-only vendor.

— **2 · Make the evidence ledger a deliverable.** Require an auditable per-action record answering who asked, who approved, what changed. It costs little when designed in and is nearly impossible to retrofit.

— **3 · Pilot on a recurring decision, with a week-one baseline.** If the workflow does not repeat on a clock, its value cannot be measured, and MIT's 95 percent finding shows where unmeasured pilots end.

— **4 · Pair open weights with an authorized frontier route.** Open-weight models inside the boundary for disconnection, frontier models through authorized regions where connectivity allows, one evaluation suite across both.

— **5 · Fund the seat, not just the licenses.** Every finding in this report points the same direction: the constraint is the embedded engineer who carries capability to acceptance. Budget for that seat explicitly.

§7 · WHAT GOOD LOOKS LIKE IN 2026

Make it measurable or do not claim it.

The market's vocabulary problem is that every vendor claims embedded, governed, and operational. The fix is to define the terms so they can be tested. We propose, and hold ourselves to, a four-part definition of done for any mission AI engagement, checkable in writing at day 90: **operational** (running in the customer's environment, under the customer's controls, on the customer's data), **measured** (a week-one value baseline and agreed measures met, or kill criteria invoked with the connectors, evaluation cases, and data map left behind), **governed** (every consequential action shows a named human approval in an evidence ledger, and no action exists outside it), and **adopted** (the owning team operates the workflow with the vendor out of the room). Any vendor, including us, can be held to this instrument; the full version is published as the Cleared FDE Standard.

---

§8 · FIVE PREDICTIONS FOR 2027

- **1 · Evidence becomes an acceptance criterion.** At least one major solicitation will require an auditable per-action record, in effect an evidence ledger, as a deliverable rather than a feature.
- **2 · The integration lead becomes a market.** As AI Integration Leads stand up across commands, demand shifts from AI builders to AI-fluent security engineers who can pass a review board. Compensation for the dual-discipline profile diverges from single-discipline peers.
- **3 · Model churn gets industrialized.** The 30-day deployment clock forces evaluation-case suites and model-swap runbooks into standard practice; "which model" becomes a weekly configuration decision, not an annual acquisition.
- **4 · The pilot graveyard gets a map.** The lessons-learned repositories GAO pushed for begin filling, and repeat-failure patterns become citable in market research. Vendors with documented acceptance records gain a structural edge.
- **5 · Readiness stays the beachhead.** The largest count of new operational AI workflows lands in readiness, maintenance, scheduling, and supply, not in headline autonomy, because that is where the data, the metric, and the mission owner already agree.

We will score these publicly in the 2027 edition, hits and misses both.

METHOD AND SOURCES

This report draws exclusively on public, unclassified sources, principally: the Department of War AI strategy, "Accelerating America's Military AI Dominance" (January 2026); DoD FY2026 budget request materials as reported in trade press (AI and autonomy line of $13.4B and its composition); GAO-26-107859, "Artificial Intelligence Acquisitions" (2026); GAO-25-107653, "Generative AI Use and Management at Federal Agencies" (2025); GAO-23-105850, "DoD Needs Department-Wide Guidance" (2023); MIT Project NANDA, "The GenAI Divide: State of AI in Business" (2025), as reported in national business press; the Stanford HAI AI Index Report 2026; OpenAI's gpt-oss release documentation; and Google Cloud's Distributed Cloud air-gapped product documentation. Engagement outcomes cited are reported by the customer organization and imply no government endorsement. Figures should be verified against the primary documents before external use; exact citations with URLs are maintained in the source annex.

DEFINITIONS USED IN THIS REPORT

— **Mission AI:** AI applied to a recurring operational decision inside a government-approved environment, measured against that mission's own metrics.

— **Cleared FDE™:** a security-cleared engineer practicing security engineering and AI engineering as one discipline, embedded with the mission owner and accountable from candidate architecture through operational acceptance. Defined publicly in the Cleared FDE Standard at federal.ai/standard.

— **Evidence ledger:** an append-only, hash-chained record of every consequential action in a governed workflow, answering who asked, who approved, and what changed. Reference schema at federal.ai/evidence-ledger.

ABOUT FEDERAL AI

Federal AI is a HUBZone small business (CAGE 7UCA4, UEI MV3AMKSSPZR1) fielding cleared forward-deployed AI engineers for U.S. federal missions, with current Department of the Air Force delivery and three deployable product lines: Optimize, Direct-To, and EnclaveDeck. The firm publishes its delivery discipline as the Cleared FDE Standard and its governance framework as the Evidence Ledger, both free to reference with attribution. This report is produced by Federal AI as a company publication from public, unclassified sources.

DISCUSS THIS REPORT

Disagree with a finding or a prediction? We would rather hear it now than in the 2027 scorecard. Write **lc@federal.ai** or book thirty minutes at **federal.ai**.