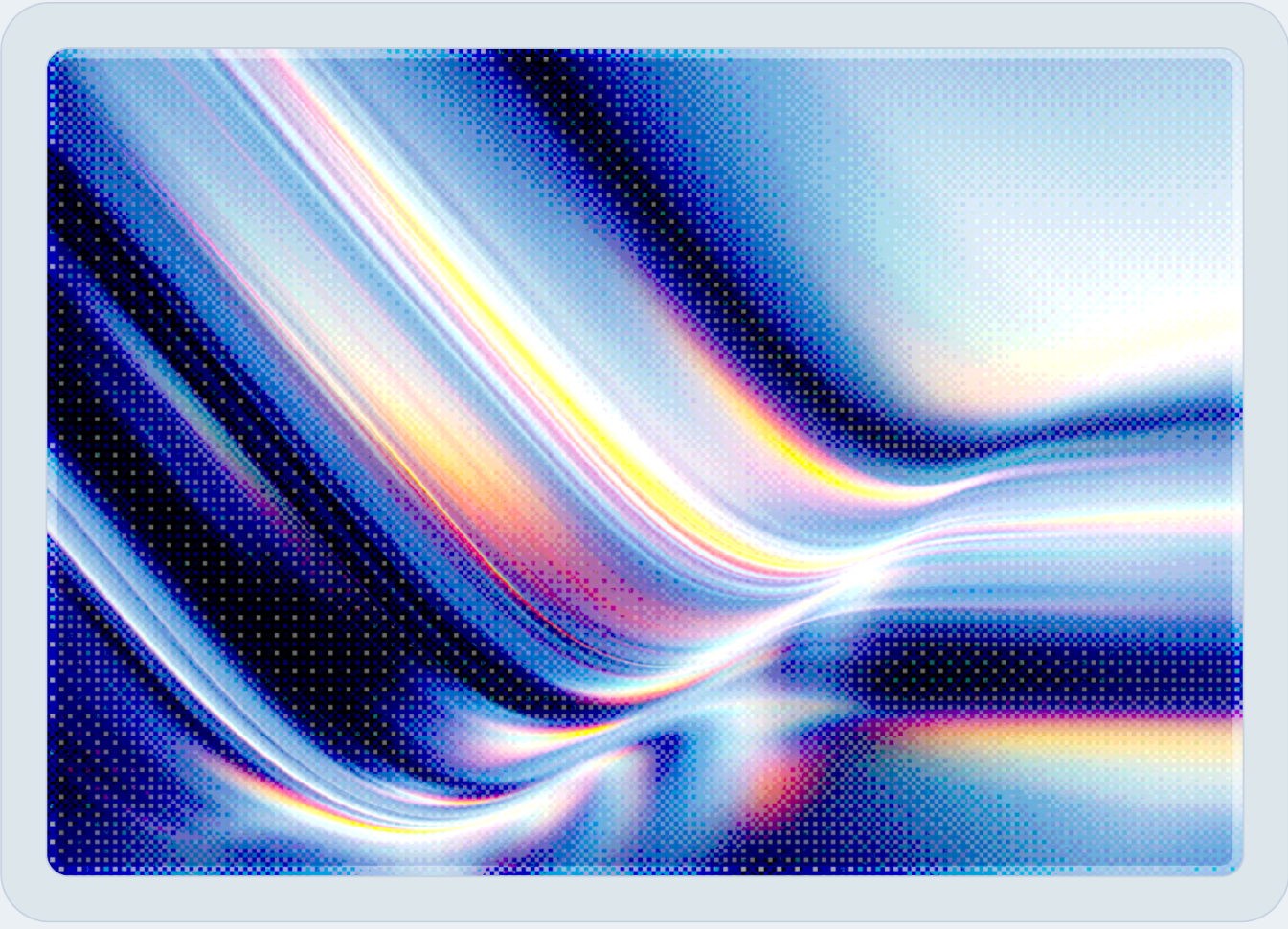
CSAI Foundation | Cloud Security Alliance

Military AI Concentration Risk and the Enterprise Pattern

100,000 Agents in Five Weeks, Classified Vendor Consolidation, and Systemic Lessons for Organizations

2026-05-04

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Table of Contents

Executive Summary 4

Introduction and Background 4

The GenAI.mil Sprint: 100,000 Agents in Five Weeks 5

Classified AI and Vendor Concentration: The Thunderforge Consortium 7

Security Implications of Rapid Agentic Deployment 8

The Enterprise Pattern: Concentration Risk Beyond Defense 9

Governance and Authorization Gaps 12

Conclusions and Recommendations 13

- Immediate Actions
- Short-Term Mitigations
- Strategic Considerations

CSA Resource Alignment 14

References 16

Executive Summary

In the spring of 2026, the U.S. Department of Defense accomplished something no organization had done publicly before: it deployed more than 103,000 semi-autonomous AI agents across its enterprise networks in fewer than five weeks [1]. Simultaneously, the department finalized deals with eight commercial AI firms for access to its most sensitive classified networks [2], while its flagship program for classified operational AI planning—Thunderforge—concentrated that mission-critical function into a consortium of three vendors: Scale AI, Anduril, and Microsoft [3].

These events are not merely a defense policy story. They constitute the clearest real-world signal yet of a structural pattern that security architects, risk officers, and enterprise technology leaders will increasingly confront: the convergence of agentic AI speed, vendor concentration, and critical-infrastructure dependency. The DoD's experience illustrates what happens when an organization moves from AI experimentation to AI-at-scale faster than governance and diversification frameworks can keep pace.

This paper examines the two interconnected concentration risks the DoD situation exposes. The first is operational concentration: when a single commercial platform becomes the engine for enterprise-wide agent creation, the security, availability, and policy posture of that platform becomes a first-order organizational risk. The second is strategic concentration: when a small number of vendors hold the keys to classified or mission-critical AI functions, political and commercial disputes—as the Anthropic episode demonstrated—can instantly impair operations at national scale [4]. Each of these risks has a direct commercial-enterprise analogue.

The Cloud Security Alliance's MAESTRO framework, AI Controls Matrix (AICM), and Zero Trust guidance provide the analytical vocabulary and control architecture needed to address both dimensions. The core recommendation of this paper is that enterprises apply diversification, interoperability, and layered authorization requirements to AI vendor relationships before agent deployments reach the scale at which remediation becomes structurally impractical.

Introduction and Background

The pace at which artificial intelligence has moved from an exploratory capability to an operational one inside the world's largest military bureaucracy has surprised even its architects. The Department of War (formerly the Department of Defense) [24] launched its enterprise AI platform, GenAI.mil, in December 2025, initially making Google Cloud's Gemini for Government available to more than three million military and civilian personnel for unclassified workflows [5]. What followed exceeded projections. Within weeks of

releasing an agent-creation tool called Agent Designer in early 2026, Pentagon users had built more than 103,000 semi-autonomous AI agents, with over 1.1 million agent sessions recorded as of mid-April [9][11]. Five of the six military branches had already designated GenAI.mil as their primary enterprise AI platform [6].

The scale of this deployment is historically significant not because it is the largest AI rollout, but because it is the fastest publicly documented agentic deployment inside any single organization, and because the organization deploying it is responsible for classified military operations, national-level intelligence, and wartime decision-making. The combination of speed, sensitivity, and agentic autonomy creates a risk profile that differs qualitatively from earlier AI deployments.

In parallel, the department's classified AI programs reached their own inflection point. The Defense Innovation Unit had awarded Scale AI a prototype contract in March 2025 for Thunderforge, the DoD's designated flagship program for AI-assisted military planning and operations [7]. Thunderforge concentrates classified operational AI into a consortium of Scale AI (as prime contractor), Anduril (integrating Scale AI's models into its Lattice mission-management system), and Microsoft (providing enterprise-grade large language model infrastructure) [3]. This program is intended to automate theater-level campaign planning, resource allocation, and wargaming at Indo-Pacific Command and European Command, handling classified information at multiple classification levels [8].

Alongside Thunderforge, the department moved in early May 2026 to expand its classified AI network access by signing agreements with eight commercial firms—Amazon Web Services, Google, Microsoft, NVIDIA, OpenAI, Reflection AI, SpaceX, and Oracle—to deploy their AI tools in Impact Level 6 and Impact Level 7 environments [2]. Notably absent from this group was Anthropic, which the Pentagon had designated as a supply chain risk to national security in February 2026 [4]. The Anthropic episode, examined in detail below, serves as the central case study of what concentration risk looks like when it becomes politically actualized rather than remaining a theoretical concern.

These events collectively define what this paper calls the enterprise pattern: the sequence by which organizations adopt AI at speed, consolidate around a small number of trusted vendors, create deep operational dependencies before governance frameworks mature, and then discover—through outages, policy disputes, or market disruption—that the dependency is more fragile than anticipated.

The GenAI.mil Sprint: 100,000 Agents in Five Weeks

Understanding the concentration risk embedded in the GenAI.mil deployment requires understanding how it was built and how users adopted it. The platform launched in December 2025 as a federated enterprise service, with Google Cloud's Gemini for Government as the first model available at Impact Level 5—the authorization tier covering the DoD's most sensitive unclassified data [5]. The CDAO (Chief Digital and

Artificial Intelligence Office) structured GenAI.mil as a platform through which commercial AI vendors would deliver tools to Pentagon users, with the CDAO managing access control and authorization rather than model development.

The adoption curve accelerated sharply in March 2026 when Agent Designer became available to GenAI.mil users. Agent Designer is a no-code and low-code tool that allows personnel to create custom AI agents by describing their function in natural language, a process sometimes described colloquially as "vibe-coding" [9]. The tool sits atop Google's Gemini models and generates agent workflows that can automate multi-step tasks—drafting after-action reports, analyzing imagery and generating summaries, parsing financial data against strategy documents, and similar staff functions that previously required hours or days of manual work. Agent Designer is explicitly designed for users without software development backgrounds, lowering the technical barrier to agent creation to near zero [10].

The consequence of combining a large existing user base, an intuitive creation interface, and a strong organizational push from leadership to accelerate AI adoption was the creation of over 103,000 agents in under five weeks [1]. More than 1.3 million DoD personnel, including military members, civilians, and contractors, used the platform in its first five months [2][25]. The agents carry formal Authorization to Operate at IL5, meaning they can process sensitive but unclassified information as part of official workflows.

From a security posture standpoint, this deployment pattern introduces several structural risks that deserve direct examination. First, the platform as a single point of failure: because GenAI.mil relies on a single underlying model family (Google Gemini) for the vast majority of its agent activity, any availability degradation, capability change, policy shift, or security incident in that model family propagates immediately across all dependent agents and workflows. A technical platform decision made years ago is now embedded in more than 100,000 automated processes. Second, the governance gap between creation and oversight: when users without programming backgrounds can create agents that take multi-step automated actions on their behalf, the question of who reviews agent behavior, what audit logging exists, and how rogue or misconfigured agents are detected becomes acute. The DoD has claimed that agents operate within the boundaries of their ATO authorization, but the mechanism for continuous validation of 103,000 agents across a platform used by 1.3 million personnel has not been publicly specified [9].

Third, and most directly relevant to enterprise organizations outside defense, the speed of adoption has created a situation in which organizational dependency outpaced organizational understanding. Agent Designer makes it simple to create agents; it does not make it simple to understand what data those agents access, what actions they will take in edge cases, or how they interact with other agents in the same environment. Security researchers have noted that in multi-agent environments, a single compromised or misconfigured agent can propagate vulnerabilities to other agents with which it interacts, with potential cascade effects across a large population of interconnected agents [12]. The DoD has created the conditions for precisely such propagation at a scale that makes manual oversight impractical.

Classified AI and Vendor Concentration: The Thunderforge Consortium

The concentration risk in unclassified AI is significant. The concentration risk in classified operational AI is categorically more consequential, because the functions at stake—theater-level military planning, campaign development, strategic assessment—are precisely the functions that adversaries most benefit from disrupting.

Thunderforge was conceived as the DoD's mechanism for accelerating AI-assisted decision-making at combatant commands. The Defense Innovation Unit awarded Scale AI the prime prototype contract in March 2025 after a competitive process [7]. The consortium structure it produced places Scale AI as the integrating prime, with Anduril integrating Scale AI's large language model capability into its Lattice platform for advanced mission modeling and simulation, and Microsoft providing enterprise LLM infrastructure and multimodal capabilities [3]. The system is intended to synthesize large volumes of operational data, generate multiple courses of action for commanders, and run AI-assisted wargames against those options—essentially automating the analytical layer of the military planning cycle.

The security stakes of concentrating this function in a three-vendor consortium are worth stating plainly. If any of the three vendors experiences a prolonged outage, a contract dispute, a regulatory action, or a security compromise, operational planning support at INDOPACOM and EUCOM could be materially degraded at precisely the moment those commands need it most. The program's design includes human oversight of all agent actions [3], which partially mitigates autonomous decision risk, but does not address the dependency risk inherent in the architecture itself.

The broader classified AI vendor landscape, which the DoD attempted to diversify in May 2026 with its eight-vendor agreement, illustrates both the administration's awareness of this problem and its current limits. Undersecretary of Defense and Chief Technology Officer Emil Michael explicitly stated that "it's irresponsible to be reliant on any one partner" when announcing the expanded vendor agreements [2]. The department's stated intent is to build a multi-vendor architecture that prevents the Joint Force from being overly dependent on a single provider. The practical challenge is that multi-vendor architectures for AI, unlike multi-vendor architectures for cloud infrastructure, do not yet have mature interoperability standards at the capability layer. Being able to switch between language models in theory is not the same as being able to maintain operational continuity in practice when one of those models is deeply embedded in a classified mission system.

The Anthropic episode makes the fragility of single-vendor or thin-vendor architectures viscerally clear. Anthropic had been, for a period, the primary commercial AI model available in DoD classified networks. When commercial negotiations broke down over the conditions under which its technology could be used—specifically, Anthropic's refusal to authorize use for fully autonomous weapons systems or domestic mass

surveillance—the Pentagon designated the company a supply chain risk in February 2026 [4]. Anthropic's legal challenges were unsuccessful in blocking the designation in the short term [13]. The result was the effective removal of an entire model family from DoD classified operations, requiring rapid substitution with alternative vendors.

This sequence—rapid adoption, deep dependency, disputed terms, forced disengagement—follows a pattern that any enterprise organization negotiating AI vendor contracts in 2026 should recognize as a prospective scenario for their own operations. The political dimensions of the Anthropic dispute were unique to the defense context, but the structural dynamics were not. An enterprise's AI vendor relationship can be disrupted by business failure, acquisition, pricing changes, terms-of-service modifications, or export control designations, all of which can occur faster than vendor-migration planning can execute. The lesson from the DoD's classified AI experience is not that eight vendors are categorically safer than three, but that the depth of operational dependency on any given vendor must be matched by a proportionate capacity to maintain operations when that vendor is unavailable.

Security Implications of Rapid Agentic Deployment

The GenAI.mil sprint and the Thunderforge architecture together represent two ends of the agentic AI risk spectrum—unclassified consumer-of-enterprise-service risk at one end, classified mission-critical dependency risk at the other—but they share a common security challenge: the governance frameworks required to manage agentic AI at scale are not yet mature, while the deployment is very much underway.

A McKinsey analysis released in 2026 found that while 74 percent of organizations planned to deploy agentic AI moderately or more extensively within two years, only 21 percent reported a mature agentic AI governance model [14]. The DoD's GenAI.mil deployment puts the department squarely in that 79 percent—organizations deploying agents at significant scale before their governance apparatus has matured to match. The security implications of this gap fall into several categories that are structurally distinct from the risks of earlier, non-agentic AI deployments.

Agentic AI systems differ from generative AI chat interfaces in a security-relevant way: they take actions in the world, not merely produce text for human review. An agent that accesses enterprise data systems, sends messages, generates reports, or triggers downstream automated workflows creates an attack surface that extends well beyond the model itself. Prompt injection—the technique by which malicious content in data the agent processes attempts to override its instructions—is more consequential when the agent is authorized to take automated actions than when it is merely answering questions [12]. A successfully injected agent inside a DoD IL5 environment that has access to planning documents, communications systems, and other agents could initiate a chain of unauthorized actions invisible to the humans nominally overseeing it.

The agent-to-agent propagation risk is particularly pronounced in large-scale deployments. Where a platform hosts more than 100,000 agents that may interact with one another—sharing data, triggering each other's workflows, or operating as components of larger automated pipelines—a single compromised or corrupted agent can serve as a vector for adversarial injection across the broader agent population [12]. This is not a hypothetical concern. Security researchers have documented multi-agent cascade scenarios in which a single malicious payload introduced into one agent's context propagated through peer agent interactions to reach a significant fraction of an interconnected agent population. In a population of 103,000 agents operating on the same enterprise platform, the blast radius of such a cascade could be operationally significant.

Identity and authorization management for agents presents a distinct challenge from human identity management. Agents must be provisioned with credentials to access the systems they automate, yet those credentials are more difficult to scope narrowly, monitor continuously, and revoke cleanly than human credentials. Recent analysis from Palo Alto Networks notes that identity weaknesses—broad permissions, inadequately scoped service accounts, credentials embedded in agent configurations—are implicated in nearly 90 percent of agentic AI security incidents [12]. DoD agents operating at IL5 under authorizations that were valid at creation time but may have drifted as the agents' functional scope evolved represent a category of risk that standard authorization review cycles are not designed to catch at scale.

Finally, the "vibe-coding" modality of agent creation—where non-technical users specify agent behavior in natural language and the platform generates the underlying automation—introduces code-quality and security risks distinct from those in traditional software development. Agents built through natural language interfaces may exhibit unexpected behaviors in edge cases that their creators did not anticipate, may have broader data access than their creators understood was being granted, and may interact with other systems in ways that violate the intent if not the letter of their authorization. The Defense Scoop report on GenAI.mil noted that the DoD claims safeguards are in place but did not specify what those safeguards are or how they scale to 103,000 agents [9]. Enterprise organizations adopting similar no-code or low-code agent creation platforms face the same opacity problem.

The Enterprise Pattern: Concentration Risk Beyond Defense

The DoD's AI deployment trajectory is being replicated, at varying speeds and without the national security stakes, across commercial enterprises globally. The structural pattern is consistent enough to warrant treating it as a category of risk in its own right rather than a collection of individual organizational decisions.

The pattern begins with rapid adoption on a single platform. Organizations discover that a commercial AI vendor's platform dramatically accelerates a set of high-value workflows. They provision access broadly across their user base, often before comprehensive governance policies are in place, because the productivity benefits are too immediate to wait for policy cycles. Usage grows faster than governance can track. As one analyst has observed, organizations find that the AI decision becomes inseparable from a much larger infrastructure commitment when a vendor's AI is deeply integrated with its cloud, its productivity suite, and its data platform—creating what has been called "ecosystem entanglement" [15].

The second phase is functional dependency. Once agents or AI-assisted workflows become part of standard operating procedures, the organization's ability to function at normal capacity depends on the continued availability of those AI capabilities. A McKinsey survey found that 47 percent of enterprise leaders reported at least one key business function that would stop working if their primary AI vendor experienced significant downtime [14]. The June 2025 global outage of OpenAI's services illustrated this concretely: customer service queues froze, automated approval workflows halted, and document-processing pipelines went dark across enterprises that had built operational dependencies on the provider. These organizations had not made a deliberate decision to make OpenAI availability a prerequisite for normal operations; the dependency had accumulated incrementally through individual team-level adoptions.

The third phase is negotiation asymmetry and term risk. As enterprise AI vendor agreements come up for renewal, organizations that have accumulated deep operational dependencies discover that they lack meaningful negotiating leverage. The vendors most deeply embedded in organizational workflows—through agent ecosystems, fine-tuned models, proprietary data connectors, and platform-specific agent architectures—are in structurally stronger negotiating positions than those whose technology can be cleanly substituted. The DoD encountered an extreme version of this dynamic in its dispute with Anthropic, where the combination of deep classified-network integration and high switching costs meant that renegotiation was politically fraught and the outcome of the dispute—a forced removal—was operationally painful regardless of which party one judges to have been correct on the substantive question.

The enterprise landscape on concentration risk reveals a widespread but underaddressed vulnerability. Industry analyses from 2026 consistently find that large majorities of enterprise technology leaders express serious concern about AI vendor dependency while acknowledging they lack the architectural preparation to execute a transition without material disruption to operations [16]. This gap—where risk awareness substantially outpaces readiness—describes a systemic problem that, in aggregate, constitutes a form of critical infrastructure vulnerability at the economy-wide level. Individual enterprises face operational disruption; in aggregate, concentration across a small number of AI model providers creates correlated failure modes across industries, supply chains, and public services.

The analogy to earlier infrastructure concentration dynamics is instructive. Cloud computing created analogous concerns about the concentration of critical workloads in a small number of providers, and those concerns led to investments in multi-cloud architectures, data portability standards, and cloud exit planning. The difference with AI vendor concentration is that the switching cost is not merely data migration and

infrastructure reconfiguration—it is the re-creation of model-specific fine-tuning, agent architectures, and human workflow adaptations that have been built around a specific model's capabilities and behaviors. Models from different providers do not have identical capabilities, and an enterprise that has built a significant agent ecosystem around one provider's model family may find that migration to an alternative requires rebuilding substantial operational logic, not merely redirecting API calls.

One mitigation pathway that has emerged is the adoption of interoperability protocols designed to decouple agent logic from specific model backends. The Model Context Protocol (MCP), originally developed by Anthropic and released as an open standard, provides a standardized interface through which agents can interact with tools and data sources across different model backends [17]. Adoption of open interoperability standards like MCP at the agent-architecture layer is the closest analogue the AI ecosystem currently has to the open standards that enabled multi-cloud portability in cloud infrastructure. Organizations that architect their agent ecosystems around MCP-compatible interfaces preserve the option to migrate model backends without rebuilding agent logic—a meaningfully different risk posture than those that build directly against proprietary APIs.

The table below summarizes the structural parallels between the DoD's experience and the enterprise pattern:

Dimension	DoD Pattern	Enterprise Pattern
Deployment speed	103,000 agents in < 5 weeks on GenAI.mil [1]	Rapid expansion of AI-assisted workflows before governance maturity
Primary vendor dependency	Google (GenAI.mil/IL5), Scale AI / Anduril / Microsoft (Thunderforge/classified) [3][5]	Single cloud provider or hyperscaler AI suite embedded in productivity stack
Governance gap	ATO granted at IL5; 103,000 agents with no publicly specified continuous oversight mechanism [9]	79% of organizations lack mature agentic AI governance models [14]
Concentration trigger event	Anthropic supply chain risk designation; forced disengagement [4]	OpenAI June 2025 outage; vendor pricing or terms changes; acquisition
Policy response	Expand to 8-vendor classified network deal; stated goal of preventing lock-in [2]	Multi-model strategies; MCP adoption; internal model hosting

Dimension	DoD Pattern	Enterprise Pattern
Interoperability investment	Limited; classified-network vendor diversity is architectural breadth, not operational portability	MCP, open APIs, model-agnostic agent frameworks emerging [17]

Governance and Authorization Gaps

The security risk of the DoD's agent deployment cannot be understood solely through the lens of technical vulnerabilities. The authorization and governance architecture surrounding large-scale agentic deployments represents an equally important risk surface, and the current state of that architecture—in defense and in enterprises—reflects the immaturity of governance practice relative to deployment practice.

Authorization to Operate in U.S. government systems is intended to certify that a system's security posture has been assessed and accepted before it handles sensitive information. IL5 ATO for the GenAI.mil platform means that the platform itself has been assessed—not each of the 103,000 agents built on it. When an individual user creates an agent using Agent Designer, that agent inherits the platform's authorization rather than receiving its own. This is operationally sensible at scale; an individual ATO assessment for each agent would be impractical. But it means that the platform's ATO is bearing a much larger burden than a traditional system ATO. The security boundaries that were assessed when the platform ATO was granted may not adequately bound the security implications of 103,000 individual agents, each accessing different data sources, automating different workflows, and potentially interacting with each other in ways that were not part of the original system assessment.

The enterprise equivalent is the practice of granting broad OAuth permissions or service account access to AI tools at the organizational level, with the expectation that individual users' use of those tools will remain within acceptable bounds. Security practitioners have documented that this expectation is frequently violated in practice, not through malicious intent but because users lack visibility into what permissions their AI tools are exercising on their behalf [12]. The combination of broad platform-level authorization and individual-level agent creation is a recipe for permission accumulation—a gradual expansion of effective access rights across an organization's AI agent population that no individual authorization decision explicitly approved.

Human oversight requirements for agentic AI are at an early stage of maturity in both defense and commercial contexts. The DoD's Thunderforge program specifies human oversight for all agent actions [3], but this requirement is qualitative rather than operational: it does not specify what forms of oversight are required for which types of actions, at what decision latency, or with what level of human understanding of

the action being authorized. An operator who approves an agent recommendation without understanding the analytical chain that produced it is technically providing oversight, but the security properties of that oversight are questionable. As agentic AI systems handle more complex and time-sensitive decisions, the practical conditions for meaningful human oversight will require explicit specification rather than general requirement.

Conclusions and Recommendations

The DoD's AI sprint of spring 2026 provides the most concrete large-scale evidence to date of what agentic AI concentration risk looks like in practice. The combination of rapid deployment, single-platform dependency, thin classified vendor architecture, and immature governance frameworks describes not an exception but a leading indicator of the conditions that will characterize enterprise AI deployments across sectors in the near term. The structural risks are manageable, but they require deliberate architectural choices that most organizations have not yet made.

Immediate Actions

Organizations deploying AI agents at significant scale should audit their current agent inventory to determine the scope of platform-level and model-level dependency. The audit should identify, at minimum, which workflows have become operationally dependent on a specific vendor, what the estimated recovery time would be if that vendor became unavailable, and what data access each agent class has been granted. For organizations that cannot answer these questions, the first priority is instrumentation and logging that makes agent behavior and data access visible before expanding agent deployments further.

Security teams should assess whether their existing Authorization to Operate or equivalent governance processes are adequate for the scale of agent deployment they are managing. Platform-level authorization does not automatically bound individual agent risk. Organizations should establish clear categories of agent action—with differentiated oversight requirements for actions that carry different risk profiles—rather than applying uniform governance to all agents regardless of what they do or what systems they access.

Short-Term Mitigations

At the architecture level, organizations should prioritize agent frameworks that implement open interoperability standards, particularly MCP, to preserve model-backend substitutability. This is not primarily a cost-reduction measure; it is a resilience measure. The goal is to ensure that the capability and continuity of agent workflows is not contingent on the continued availability of a specific vendor's model infrastructure.

Vendor agreements for AI services should explicitly address continuity provisions: data portability, API stability commitments, advance notice for capability changes, and the conditions under which the vendor can alter or terminate service. The DoD's experience with Anthropic illustrates that vendor agreements that do not address edge conditions become subject to improvised resolution under pressure—and that the outcomes of those improvisations may not serve organizational interests.

Multi-model testing, even at modest scale, should begin before it is needed operationally. Organizations that have evaluated alternative model providers for their critical workflows can execute transitions more rapidly when forced to than organizations encountering alternatives for the first time under pressure.

Strategic Considerations

The enterprise pattern of AI vendor concentration is developing faster than regulatory or standards frameworks are developing to address it. Organizations that invest now in governance maturity, architectural flexibility, and vendor risk assessment will be better positioned both operationally and reputationally when concentration-related disruptions occur in their sector. Conversely, organizations that defer these investments until a concentration event forces the issue will find the remediation far more costly.

The DoD's stated goal of building an AI architecture that prevents over-reliance on any single vendor is the right strategic orientation [2]. The practical challenge is that this goal must be pursued architecturally, not only contractually. Having agreements with eight vendors for classified network access is architecturally different from having eight interoperable, operationally equivalent options. Enterprises should make the same distinction in their own multi-vendor AI strategies: contractual diversification without operational portability leaves organizations exposed to the same concentration risks under different terminology.

CSA recommends that organizations adopt a formal AI vendor risk assessment process, comparable in rigor to third-party risk management programs for other critical infrastructure dependencies, with explicit concentration thresholds that trigger diversification requirements. As agentic AI becomes embedded in more workflows and at greater depth, the cost of remediating concentration risk will rise. The time to build resilient architectures is before 100,000 agents are deployed, not after.

CSA Resource Alignment

The risks described in this paper map directly to frameworks and guidance published by the Cloud Security Alliance, providing organizations with structured analytical tools to assess and address their own exposure.

MAESTRO (Multi-Agent Environment, Security, Threat, Risk, and Outcome) is CSA's layered threat modeling framework for agentic AI systems, introduced in February 2025 and expanded with real-world application guidance in 2026 [18][19]. MAESTRO's seven-layer architecture—from Foundation Models through Agent Ecosystem—captures precisely the attack surfaces that the DoD's deployment exposes: supply chain risks in model layers, orchestration vulnerabilities in agent management layers, and cross-layer cascade effects that begin at the model level and propagate through to operational impacts. Organizations seeking to assess the security posture of their agent deployments should apply MAESTRO as the primary threat modeling framework, with particular attention to the Agent Ecosystem layer, where vendor concentration risk is most architecturally visible.

CSA's AI Controls Matrix (AICM) provides a control framework mapped across the AI shared responsibility model, covering supply chain security, model governance, data security, and orchestration controls [20]. The AICM's supply chain security domain is directly applicable to vendor concentration risk: it provides control language for assessing vendor dependency, continuity planning, and architectural substitutability. Organizations deploying agents should use the AICM as the basis for their agent governance programs, mapping existing controls against AICM requirements to identify gaps.

CSA's Zero Trust guidance applies to the identity and access management dimensions of agentic AI described in this paper [21]. The principle of least-privilege access, continuous verification, and explicit authorization—core Zero Trust tenets—are directly applicable to agent credential management, data access scoping, and human oversight requirements. Treating agents as untrusted principals that must earn authorization for each action, rather than trusted actors with broad inherited permissions, is the Zero Trust posture appropriate to the current threat environment.

CSA's Cloud Controls Matrix (CCM) provides governance controls for cloud infrastructure dependencies, including availability requirements, business continuity provisions, and supply chain risk management [22]. As AI services are delivered primarily through cloud infrastructure, CCM controls for cloud vendor management are directly applicable to AI vendor risk, particularly in the areas of exit planning, data portability, and incident response coordination with vendors.

Finally, CSA's work on **AI Organizational Responsibilities** addresses the governance structures organizations need to manage AI risk at the leadership level—board-level oversight, risk ownership, and accountability frameworks that ensure AI vendor concentration is treated as a material operational risk rather than a technical procurement matter [23].

References

[1] DefenseScoop. "Pentagon uses GenAI.mil to create 100K agents." DefenseScoop, April 23, 2026.

[2] DefenseScoop. "DOD expands its classified AI work with 8 companies — excluding Anthropic — amid ongoing dispute." DefenseScoop, May 1, 2026.

[3] Scale AI. "Introducing Thunderforge: AI for American Defense." Scale AI Blog, March 2025.

[4] NBC News. "Anthropic says the Pentagon has declared it a national security risk." NBC News, 2026.

[5] Google Cloud. "Gemini for Government: Build custom AI agents for unclassified work on GenAI.mil." Google Cloud Blog, 2026.

[6] DefenseScoop. "5 out of 6 military branches have elevated GenAI.mil as their go-to enterprise AI platform." DefenseScoop, February 2, 2026.

[7] Defense Innovation Unit. "DIU's Thunderforge Project to Integrate Commercial AI-Powered Decision-Making." DIU, March 2025.

[8] Breaking Defense. "AI for war plans: Pentagon innovation shop taps Scale AI to build 'Thunderforge' prototype." Breaking Defense, March 5, 2025.

[9] Breaking Defense. "Pentagon workers vibe-code 100,000 AI 'agents' to use on unclassified networks." Breaking Defense, April 2026.

[10] DefenseScoop. "Pentagon says employees can create their own 'custom AI assistants' with new tech." DefenseScoop, March 10, 2026.

[11] The Defense Post. "Pentagon Deploys 100K+ AI Agents to Drive Daily Workflows." The Defense Post, April 27, 2026.

[12] Palo Alto Networks. "Agentic AI vs. AI Agents: Differences, Risks & Security." Palo Alto Networks Cyberpedia, 2025.

[13] CNBC. "Anthropic loses appeals court bid to temporarily block DOD ruling." CNBC, April 8, 2026.

[14] McKinsey & Company. "Deploying agentic AI with safety and security: A playbook for technology leaders." McKinsey, 2026.

[15] Kai Waehner. "Enterprise Agentic AI Landscape 2026: Trust, Flexibility, and Vendor Lock-in." Kai Waehner Blog, April 6, 2026.

[16] Ability.ai. "AI vendor lock-in risks: the operational crisis CEOs must address." Ability.ai Blog, 2026.

[17] MindStudio. "What Is the Transitional Lock-In Risk in AI Agent Infrastructure?." MindStudio Blog, 2026.

[18] Cloud Security Alliance. "Agentic AI Threat Modeling Framework: MAESTRO." CSA Blog, February 6, 2025.

[19] Cloud Security Alliance. "Applying MAESTRO to Real-World Agentic AI Threat Models." CSA Blog, February 11, 2026.

[20] Cloud Security Alliance. "AICM Implementation & Auditing Guidelines." Cloud Security Alliance, 2025.

[21] Cloud Security Alliance. "Zero Trust Guidance for Critical Infrastructure." Cloud Security Alliance, 2025.

[22] Cloud Security Alliance. "Cloud Controls Matrix (CCM)." Cloud Security Alliance, 2025.

[23] Cloud Security Alliance. "AI Organizational Responsibilities." Cloud Security Alliance, 2025.

[24] U.S. Department of War. "War Department Launches AI Acceleration Strategy." Department of War Press Release, 2026.

[25] Defense One. "Pentagon adds Google's latest model to GenAI.mil as usage soars." Defense One, April 2026.