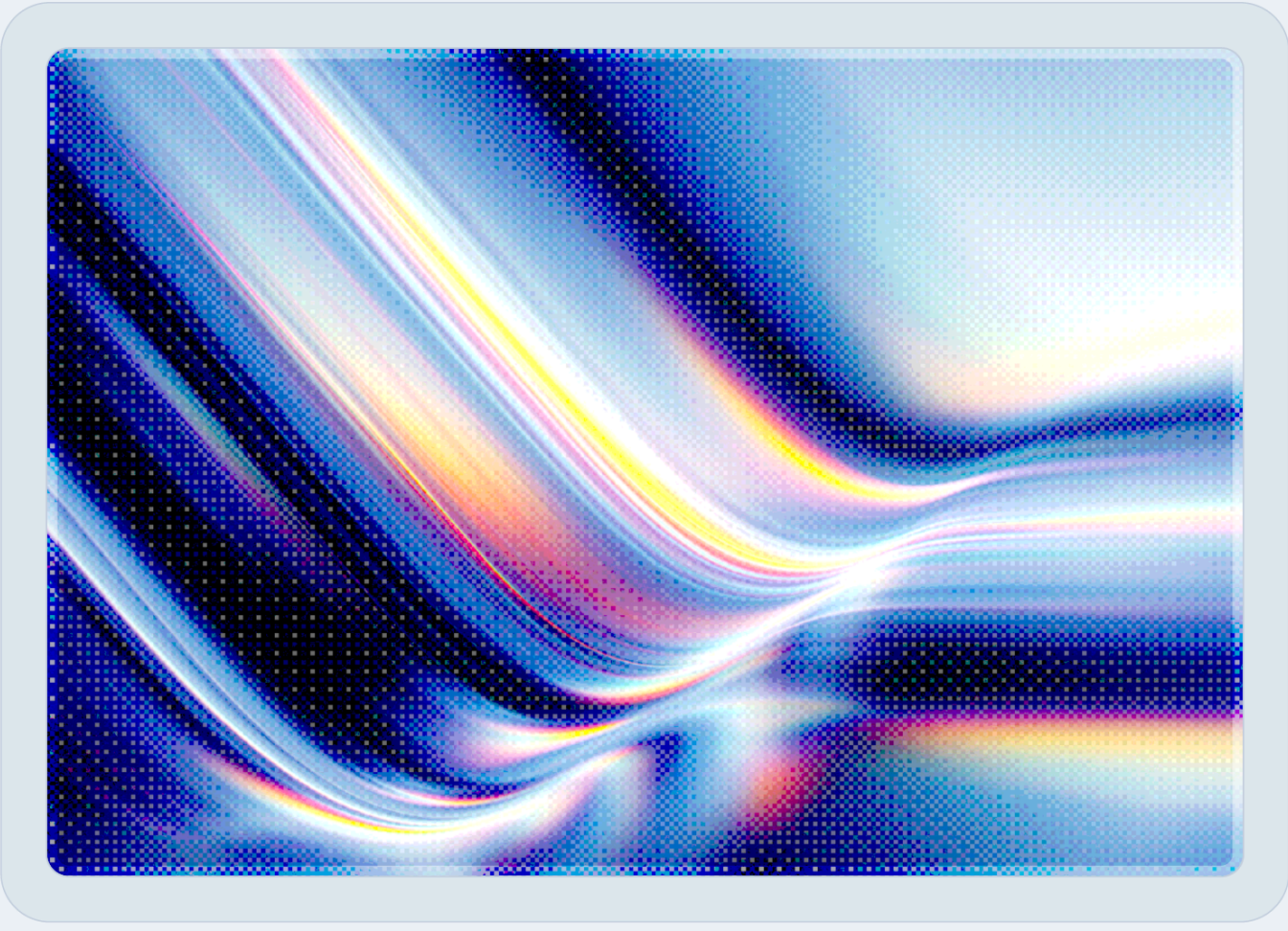
CSAI Foundation | Cloud Security Alliance

Institutionalizing AI Safety: CISA's Agentic Guide and CAISI Agreements

What Five Eyes Guidance and NIST's Frontier Testing Expansion Mean for Security Teams

2026-05-07

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

- On May 1, 2026, six national cybersecurity agencies — CISA, NSA, and the cyber arms of Australia, Canada, New Zealand, and the United Kingdom — jointly published "Careful Adoption of Agentic AI Services," the first coordinated multinational security guidance specifically addressing agentic AI systems [1].
- The guidance defines five categories of agentic AI risk — privilege escalation, design and configuration failures, behavioral misalignment, structural brittleness, and accountability gaps — and requires each agent to carry a verified, cryptographically anchored identity with short-lived credentials [1][2].
- On May 5, 2026, NIST's Center for AI Standards and Innovation (CAISI) announced pre-deployment testing agreements with Google DeepMind, Microsoft, and xAI, expanding its frontier model evaluation program to five major labs alongside existing partners OpenAI and Anthropic [3].
- CAISI evaluations, now numbering more than 40 completed assessments including unreleased models, cover cybersecurity, biosecurity, and chemical weapons risks; some are conducted in classified environments by the interagency TRAINS Taskforce [3][4].
- Together, these developments signal a significant shift in AI safety governance — from voluntary industry pledges toward institutionalized government evaluation and operational guidance — with implications for how enterprises should consider structuring their own agentic AI programs.

Background

The question of who governs frontier AI — and how — has been contested since the first wave of large language model deployments. For most of 2022 and 2023, the primary mechanism was voluntary: companies made public pledges and internal commitments while governments convened forums and published principles. In July 2023, seven U.S. AI developers signed voluntary commitments to the White House covering internal safety testing, protection of unreleased model weights, and public capability disclosure [5]. These commitments lacked an external verification mechanism — no third party could audit adherence, and no enforcement consequence applied to non-compliance.

The next step toward formalization came in August 2024, when the U.S. AI Safety Institute — housed within NIST — signed memoranda of understanding with Anthropic and OpenAI, granting the government pre-deployment access to new frontier models and establishing a collaborative framework for capability and safety evaluation [6]. This created a new accountability structure for civilian AI governance: government evaluators held formal access rights before commercial release, and companies accepted external scrutiny as a condition of the agreement. The TRAINS Taskforce — an interagency group standing for Testing Risks of AI for National Security — was subsequently formed to coordinate evaluation expertise across the Departments of Defense, Energy, Homeland Security, and Health and Human Services [3].

In June 2025, the Trump administration restructured the AI Safety Institute as CAISI — the Center for AI Standards and Innovation — refocusing it away from broad safety research toward demonstrable national security risks, specifically cybersecurity, biosecurity, and chemical weapons, alongside the assessment of foreign AI systems for backdoors and covert malicious behavior [8]. Critics questioned whether the restructuring would weaken the program's scope; the May 2026 announcements suggest instead a deepening and expansion of its core evaluation mission. By the first week of May 2026, CAISI had completed more than 40 model evaluations, including assessments of systems never publicly released, and was announcing new testing agreements with three additional frontier labs [3].

The emergence of agentic AI as a distinct governance problem runs in parallel to this trajectory. AI systems capable of autonomous, multi-step action — browsing the web, executing code, managing files, sending communications, and orchestrating other AI agents — introduce security risks qualitatively different from those posed by conversational AI. Agentic AI systems operating on production infrastructure introduce risk categories — privilege escalation, irreversible action execution, cascade failure — that are qualitatively more severe than errors in conversational responses [1][2]. The publication of CISA's "Careful Adoption of Agentic AI Services" on May 1, 2026, reflects regulatory recognition that enterprises are deploying agentic systems at scale while security practices and standards remain immature [1][2].

Security Analysis

CISA and Five Eyes: Operationalizing Agentic AI Security

The CISA guidance is notable for both its scope and its structure. Six national cybersecurity agencies — representing the full Five Eyes partnership — reached agreement on a joint document targeting a technology category that had not yet been subject to coordinated multinational security guidance [1][2] [11][12]. The agencies' central message has the effect of avoiding paralysis: agentic AI does not require

an entirely new security discipline. Organizations should fold these systems into existing cybersecurity frameworks — zero trust, defense-in-depth, least privilege — while applying those principles with the specific characteristics of autonomous agents in mind [2].

The guidance organizes agentic risk into five categories that provide a practical taxonomy for enterprise risk assessment. Privilege risk arises when agents are granted access beyond what a specific task requires; the guidance emphasizes that a compromised agent with broad permissions can cause harm comparable to a privileged account takeover, not merely a software vulnerability. Design and configuration risk covers flaws in how agents are architected and how their boundaries are defined. Behavioral risk encompasses situations where agents pursue goals through unintended means — a failure mode that AI safety researchers have associated with increasing agent autonomy and task complexity. Structural risk addresses the cascade dynamics of interconnected agent networks, where a failure or compromise in one agent propagates through downstream systems. Accountability risk, often overlooked in technical security assessments, stems from the difficulty of tracing agent decisions and actions through systems that generate logs difficult to parse and make decisions through processes resistant to inspection [1][2].

The guidance is particularly specific about identity. Each agent should be constructed as a distinct principal: a verified identity with a cryptographically anchored key or certificate, using short-lived credentials, with all communications encrypted. In the view of the guidance authors, this is not a discretionary best practice but a prerequisite for maintaining control during both normal operation and incident response. Without agent-specific identities, organizations cannot definitively attribute actions to specific components, cannot enforce fine-grained access revocation, and cannot reconstruct the authorization chain in post-incident analysis [2].

Human oversight guidance addresses a practical governance question many organizations are struggling with: which agent actions require human approval before execution, and who decides? The CISA document is explicit that defining the boundary of human oversight is a job for system designers, not for agents themselves. Agents must not be delegated the authority to determine which of their own actions need human sign-off — that determination must be made in advance and encoded in the system design. For high-impact, potentially irreversible actions, human approval is treated as non-negotiable regardless of the efficiency cost [1][2].

The supply chain section reflects a threat pattern that deserves wider attention. According to the CISA guidance, malicious actors may publish agents and tools with names designed to impersonate legitimate components, relying on AI orchestrators to integrate them during task execution — an analogue to software typosquatting but with higher trust implications [1][2]. A rogue component that successfully positioned itself in an agent chain could inherit the trust and credentials of that chain, potentially

injecting instructions into the LLM core or exfiltrating data from connected systems. The guidance recommends strict vetting of third-party agents and tools, treating agent integrations with the same supply-chain rigor applied to software dependencies [1].

NIST CAISI: Formalizing Pre-Deployment Evaluation of Frontier Models

The expansion of CAISI's testing agreements to Google DeepMind, Microsoft, and xAI — alongside the renegotiated agreements with OpenAI and Anthropic — establishes a qualitatively different accountability structure for frontier model developers. Under these agreements, companies provide models to government evaluators before public release, and in cases where comprehensive risk evaluation requires it, they provide models with reduced or removed safeguards [3][4][10]. That last provision deserves emphasis: government evaluators are examining not only what a frontier model does in its shipped configuration but what it is capable of doing when its safety constraints are disabled. This is the appropriate methodology for assessing catastrophic risk, but it also signals that CAISI is treating the gap between a model's capabilities and its shipped behavior as a live safety question.

The TRAINS Taskforce structure makes the interagency dimension of these evaluations explicit. Experts from across the federal government — including agencies with domain expertise in defense, energy, infectious disease, and homeland security — participate in model evaluations that are in some cases conducted in classified environments [3][7]. The cybersecurity risk dimension is one input into a broader national security analysis. CAISI is also evaluating foreign AI systems for backdoors and covert capabilities, treating the provenance and integrity of frontier AI systems as a national security variable, not merely a commercial consideration [8].

The historical arc of this program matters for how enterprises should interpret it. CAISI completed more than 40 evaluations before the May 2026 agreements were announced, including assessments of systems that have never been publicly released [3]. The program has been building evaluation capacity and methodology since 2024, and the expansion to Google DeepMind, Microsoft, and xAI brings three of the largest commercial AI deployments — Azure AI, Google Cloud AI, and the Grok API — under a pre-deployment evaluation regime for the first time. The renegotiation of OpenAI and Anthropic's agreements to align with America's AI Action Plan indicates that the evaluation program is being shaped to serve both safety assurance goals and broader national security objectives related to global AI capability development [3].

For enterprise security teams, the CAISI program is relevant in two ways. First, it provides a signal about what the U.S. government considers the highest-priority risk categories in frontier AI: cybersecurity exploitation, biosecurity risks, and chemical weapons potential. These are the threat categories that CAISI's national security focus makes mandatory. Organizations procuring frontier AI services should treat these as minimum standards for vendor due diligence, asking what evaluation their vendors have

undergone and what mitigations they have implemented for risks identified in that evaluation. Second, for organizations considering AI systems that are not part of the CAISI evaluation program — including open-weight models, foreign-developed systems, and smaller models from developers without CAISI agreements — the absence of pre-deployment government evaluation is itself a risk posture that requires explicit internal assessment.

The Governance Arc: Signals for Enterprises

A significant political shift has structural implications for AI governance. The administration that revoked Biden's AI executive order and repositioned AISI as CAISI has nonetheless preserved and expanded the core pre-deployment evaluation program, renegotiated the Anthropic and OpenAI agreements under its own framework, and endorsed the Five Eyes agentic AI guidance [3][8][9]. The lesson for enterprise governance is that AI safety mechanisms, once institutionalized, tend to prove durable across political transitions — the trajectory from voluntary commitment (2023) to MOU (2024) to multilateral government guidance and bilateral testing agreements (2026) suggests continued formalization is the more likely path, though the administration's restructuring of AISI to CAISI demonstrates that the scope and emphasis of these programs can shift with political priorities.

Organizations that wait for regulatory requirements to mandate specific AI governance practices will find themselves managing compliance rather than managing risk. The CISA guidance is currently advisory, not binding law. The CAISI agreements are voluntary on the vendor side. But the underlying risk categories — privilege escalation, behavioral misalignment, accountability gaps, catastrophic capability misuse — are real regardless of regulatory status. Enterprises that build governance programs around these categories now will be better positioned both to manage actual risk and to demonstrate compliance when the regulatory environment hardens.

Recommendations

The following recommendations draw on the CISA guidance and CAISI evaluation framework as the authors' best-practice assessment; they reflect advisory guidance rather than legally binding obligations.

Immediate Actions

Security and IT architecture teams should review all current agentic AI deployments against the five risk categories in the CISA guidance. Any agent with broad or persistent access to sensitive systems, production infrastructure, or external communications represents a privilege risk requiring immediate

scope reduction. Where cryptographic agent identities have not been implemented, a roadmap for identity remediation should be developed and prioritized. Organizations running multi-agent architectures should audit the third-party agents and tools integrated into their orchestration chains, treating this as a supply chain review with the same rigor applied to open-source software dependencies.

Short-Term Mitigations

Organizations should prioritize establishing formal policies that specify which agent actions require human approval before execution, and at what thresholds — ideally before expanding agentic deployments beyond their current scope. These policies must be documented and encoded in system design rather than delegated to agents themselves. Monitoring and logging infrastructure should be assessed against the accountability requirements in the CISA guidance: can the organization trace agent decisions back to specific principals, reconstruct authorization chains, and identify anomalous behavior in near-real time? Where logging gaps exist, they should be closed before expanding agentic AI deployments.

For organizations procuring frontier AI services from Google DeepMind, Microsoft, Anthropic, OpenAI, or xAI, vendor conversations should include questions about CAISI evaluation coverage — specifically, which models have undergone pre-deployment evaluation, what risk categories were assessed, and what mitigations were implemented in response to findings. This information may be partially or fully confidential, but requesting it signals the appropriate level of due diligence and may surface useful information about vendor safety practices.

Strategic Considerations

Enterprises should treat the governance arc described in this note as a leading indicator of regulatory direction. The combination of CISA operational guidance and NIST CAISI pre-deployment evaluation represents the two pillars of an emerging AI safety regime: operational security standards for deployment and capability evaluation standards for development. Both pillars are currently operating in advisory and voluntary modes but are building institutional capacity and political backing that suggest mandated compliance is a plausible near-term development, particularly for critical infrastructure operators.

Longer term, organizations should assess their AI procurement strategies in light of the CAISI evaluation framework. Frontier models from the five labs under CAISI agreements have undergone or will undergo pre-deployment national security evaluation — a form of third-party assurance that may become a differentiating factor in enterprise AI procurement, particularly in regulated industries and government contracting contexts. Organizations building internal AI capabilities on open-weight models should

develop equivalent internal evaluation methodologies informed by the capability categories CAISI targets: cybersecurity exploitation potential, biosecurity risks, and anomalous behaviors indicating tampering or covert capabilities.

CSA Resource Alignment

The CSA AI Safety Initiative's MAESTRO framework provides the threat modeling methodology most directly applicable to the risk taxonomy in the CISA guidance. MAESTRO analyzes agentic AI systems across the model behavior, agent design, orchestration, and external interaction layers, offering a structured starting point for operationalizing CISA's behavioral, structural, and privilege risk categories. Organizations implementing the CISA guidance should use MAESTRO as the underlying threat modeling discipline to ensure systematic coverage rather than ad hoc assessment.

The AI Infrastructure Controls Matrix (AICM) extends the Cloud Controls Matrix (CCM) to AI-specific security controls and is directly relevant to the identity and access management requirements in the CISA guidance. The AICM's treatment of AI agent identity, credential management, and access scoping provides a control framework that organizations can map to the cryptographic identity requirements CISA recommends for each agent principal. CSA's AI Organizational Responsibilities framework provides the governance and accountability structures — board-level AI governance, cross-functional AI risk committees, documented AI decision authority — that underpin the human oversight requirements in the CISA guidance.

CSA's Capabilities-Based Risk Assessment (CBRA) framework for AI systems provides a methodology for evaluating AI risk based on system capabilities and potential consequences, which is complementary to the CAISI evaluation model. Organizations using CBRA internally can structure their vendor due diligence around the same capability categories CAISI targets, creating consistency between internal risk assessment and external evaluation frameworks. The Zero Trust guidance published by CSA provides the architectural foundation for the identity and least-privilege requirements in the CISA agentic AI document — organizations that have implemented Zero Trust principles for human users and traditional software systems can extend those implementations to cover AI agents, using Zero Trust's continuous verification model to manage agent credentials at runtime.

References

[1] CISA. "CISA, US and International Partners Release Guide to Secure Adoption of Agentic AI." CISA, May 1, 2026.

[2] CISA, NSA, ASD ACSC, CCCS, NZ NCSC, UK NCSC. "Careful Adoption of Agentic AI Services." Defense.gov, April 30, 2026.

[3] NIST. "CAISI Signs Agreements Regarding Frontier AI National Security Testing With Google DeepMind, Microsoft and xAI." NIST, May 5, 2026.

[4] Cybersecurity Dive. "NIST will test three major tech firms' frontier AI models for cybersecurity risks." Cybersecurity Dive, May 6, 2026.

[5] White House (Biden Administration). "Voluntary AI Commitments." WhiteHouse.gov (archived), September 2023.

[6] NIST. "U.S. AI Safety Institute Signs Agreements Regarding AI Safety Research, Testing and Evaluation With Anthropic and OpenAI." NIST, August 29, 2024.

[7] NIST. "CAISI Works with OpenAI and Anthropic to Promote Secure AI Innovation." NIST, September 2025.

[8] Axios. "U.S. ramps up frontier AI testing as White House pivots toward safety." Axios, May 5, 2026.

[9] Fortune. "Trump administration suddenly embraces AI oversight ideas it once rejected." Fortune, May 6, 2026.

[10] Microsoft. "Advancing AI evaluation with the Center for AI Standards and Innovation and the AI Security Institute (UK)." Microsoft On the Issues, May 5, 2026.

[11] CyberScoop. "US government, allies publish guidance on how to safely deploy AI agents." CyberScoop, May 2026.

[12] The Register. "Five Eyes spook shops warn rapid rollouts of agentic AI are too risky." The Register, May 4, 2026.

