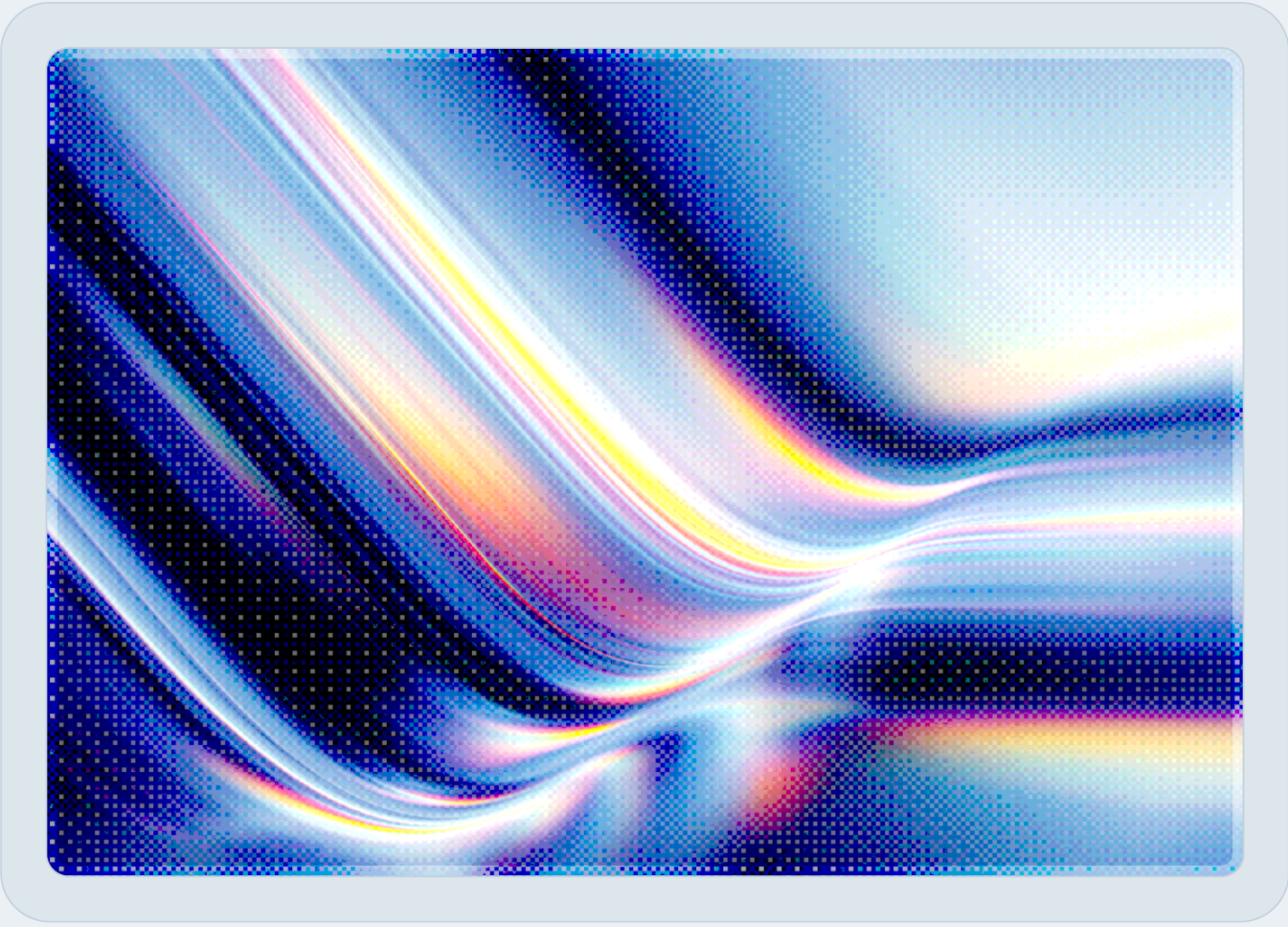
CSAI Foundation | Cloud Security Alliance

NIST CAISI Frontier AI Pre-Deployment Testing

Security Implications and Enterprise Compliance Alignment

2026-05-07

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On May 5, 2026, NIST's Center for AI Standards and Innovation (CAISI) announced voluntary pre-deployment evaluation agreements with Google DeepMind, Microsoft, and xAI — extending a program that already covered OpenAI and Anthropic. The five developers that have now entered CAISI evaluation agreements collectively include the most widely deployed U.S.-developed frontier AI systems available to enterprise customers [1].
- CAISI, formerly the U.S. AI Safety Institute (AISI), was renamed and repositioned in June 2025 with a refocused mission emphasizing AI security and national security over broad AI safety. The scope of evaluations targets demonstrable risks — cybersecurity, biosecurity, chemical and biological threats, foreign AI systems, and covert model behavior such as backdoors [2].
- Pre-deployment evaluations involve access to model versions with reduced or removed safety guardrails, allowing government evaluators to probe raw capabilities. Testing occurs in classified environments with participation from the TRAINS Taskforce (Testing Risks of AI for National Security), an interagency body spanning the Departments of Defense, Energy, Homeland Security, and Health and Human Services [3].
- Enterprise security teams procuring frontier AI products should treat CAISI evaluation coverage as one input in AI procurement due diligence — not as a certification or guarantee of safety. The program is voluntary, its findings are not public, and it addresses national security risks rather than enterprise-specific threat models.
- CSA's AI Controls Matrix (AICM) and associated MAESTRO threat modeling framework provide actionable governance structures that enterprises can apply now, independent of whether their AI vendor has undergone CAISI evaluation.

Background

The U.S. government has pursued structured pre-release access to frontier AI systems for evaluation since August 2024, when the then-U.S. AI Safety Institute signed the first such agreements with OpenAI and Anthropic [4]. Those Memoranda of Understanding established the principle that a federal agency

could receive access to major AI models both prior to and following public release, enabling collaborative research on evaluation methodology, capability assessment, and safety risk mitigation. At that stage, the institute operated under the Biden administration's Executive Order 14110 framework, with a broad mandate covering long-term safety and societal risk.

In June 2025, Secretary of Commerce Howard Lutnick reorganized and renamed the institute as the Center for AI Standards and Innovation (CAISI) [2]. The mission shifted from an emphasis on broadly defined AI safety toward a more targeted focus on AI security – specifically, on demonstrable near-term national security risks. Critics characterized the change as a pivot from long-term risk mitigation toward prioritizing U.S. AI competitiveness and security [2]. Officials at NIST and the Department of Commerce framed the refocused mandate as a more practical and immediate contribution, directing government resources toward harms that can be concretely measured and tested [3].

In September 2025, CAISI renewed its working agreements with OpenAI and Anthropic under the new mandate. Both companies published accounts of specific security improvements resulting from CAISI evaluations – including joint red-teaming of biological misuse safeguards and vulnerability testing of Anthropic's Constitutional Classifiers defense system against jailbreak attempts [5]. The September 2025 NIST announcement marked among the first substantive public accounts of concrete outcomes attributable to government-industry AI evaluation collaboration under this program.

The expansion announced on May 5, 2026 brings Google DeepMind, Microsoft, and xAI into the program under substantially similar terms [1, 10]. The five organizations that have signed CAISI evaluation agreements collectively develop the frontier AI models most widely available to enterprise customers. The program has completed more than 40 evaluations to date, including assessments of state-of-the-art systems that remain unreleased [1].

Security Analysis

Scope of Evaluation and Testing Methodology

CAISI's evaluation agreements cover both pre-deployment (before public release) and post-deployment testing, as well as targeted collaborative research. The scope is not generic AI safety auditing – the program explicitly targets national security risk domains: cybersecurity offense and defense capabilities, biosecurity and chemical threat facilitation, conventional and nuclear/radiological military applications, assessments of foreign AI systems, and identification of backdoors or other covert malicious behavior embedded in model weights [3].

AI developers typically provide CAISI with model versions in which safety guardrails have been reduced or removed [1], allowing government researchers to probe underlying model capabilities without interference from production-level safety mitigations that might suppress or obscure dangerous behaviors. The approach is conceptually analogous to red-teaming an application after disabling WAF rules — the goal is to understand what the system is capable of under adversarial conditions, not what it does under normal operating constraints.

Testing occurs in classified environments when necessary, reflecting the sensitivity of both the capabilities being assessed and the evaluation findings themselves [1]. This classification means that specific evaluation outcomes for individual models are not publicly available, which has implications for how enterprises can and cannot use this program in their own governance processes.

The TRAINS Taskforce

The interagency TRAINS Taskforce (Testing Risks of AI for National Security) was established in November 2024 and serves as the collaborative mechanism through which domain expertise from across the federal government is brought to bear on AI evaluations [3]. Its membership spans the Department of Defense — including the Chief Digital and AI Officer's office and the National Security Agency — the Department of Energy along with ten of its National Laboratories, the Department of Homeland Security and CISA, and the National Institutes of Health.

The institutional design of TRAINS reflects the multidisciplinary nature of frontier AI risk. Capability to assist in engineering a novel pathogen is not a question that a computer scientist can fully evaluate alone; it requires biosecurity domain knowledge and laboratory infrastructure. Similarly, assessing whether a model lowers the barrier for attacking industrial control systems requires input from practitioners who understand both offensive AI capabilities and operational technology security. TRAINS assembles these specialized perspectives into a coordinated evaluation structure.

The taskforce also co-develops AI evaluation methodology. This research function — producing shared benchmarks, datasets, and assessment workflows for AI security and robustness evaluation — is the component of CAISI's work most likely to produce publicly usable outputs over time, even when specific model findings remain classified [3].

The International Dimension

CAISI does not operate in isolation. The U.K.'s AI Security Institute (AISI, not to be confused with the former U.S. AISI) operates in parallel, and there is documented collaboration between the two bodies. Microsoft's May 2026 agreement explicitly names both CAISI and the UK AISI [6]. In September 2025, joint evaluation work between the U.S. CAISI and UK AISI produced security improvements in Anthropic's

Constitutional Classifiers prior to Claude model releases [5]. This bilateral coordination reflects a shared recognition that frontier AI capabilities are relevant to allied national security interests and that unilateral evaluation has gaps.

Enterprise Implications: What This Program Is and Is Not

Security professionals supporting enterprise AI procurement and governance should approach CAISI's evaluation program with calibrated expectations. It represents a concrete development in AI oversight infrastructure, but several characteristics limit its direct utility as an enterprise compliance signal.

The program is voluntary. Participation is governed by bilateral agreements that each AI developer negotiated with the government; there is no legal mandate requiring pre-deployment testing, and no mechanism compels disclosure of evaluation results to the public or to customers [1]. The absence of mandatory disclosure means that an enterprise evaluating two AI systems cannot determine from public information whether either was evaluated by CAISI or what the evaluation found.

The program's threat model is national security-centric, not enterprise-centric. CAISI evaluates risks relevant to nation-state-level harm: weapons facilitation, foreign intelligence operations, catastrophic biological or chemical threats. These are important risks, but they are generally not the primary AI threat surface for most enterprise security teams, whose concerns involve data exfiltration, prompt injection, model manipulation, unauthorized privilege escalation in agentic systems, and model behavior in regulated contexts.

Furthermore, a model that has undergone CAISI evaluation and received no adverse findings in the national security domains of interest is not thereby validated as safe for deployment in a healthcare, financial services, or critical infrastructure context. Enterprise-grade AI security governance should include independent assessment against business-specific threat models; CAISI participation alone cannot substitute for evaluation against enterprise-relevant threat surfaces.

That said, the program creates structural benefits for the enterprise security community. It normalizes the expectation that AI developers grant independent parties pre-release access for evaluation, which strengthens the case for enterprise customers demanding similar access — or at minimum, substantive third-party audit reports — during procurement. It also signals that government regulators and agencies are investing in AI evaluation infrastructure, which will likely accelerate the development of sector-specific AI regulations in financial services, healthcare, and critical infrastructure over the coming years.

Recommendations

Immediate Actions

Enterprise security and procurement teams should incorporate CAISI participation status into AI vendor questionnaires, framing it as a positive governance indicator rather than a binary qualification. Vendors participating in the CAISI program have committed to a higher level of transparency with at least one major government evaluator, which is a useful data point about vendor governance culture.

Organizations that develop or fine-tune their own AI models — rather than using them as black-box APIs — should benchmark their internal evaluation programs against what is now known about CAISI methodology. Key questions include: Does your evaluation program include testing with guardrails disabled or reduced? Does it cover CBRN-adjacent capability uplift risks that your sector regulator may eventually require you to address? Are your evaluation datasets and workflows reproducible and auditable?

Short-Term Mitigations

Security teams should not treat CAISI participation as a substitute for enterprise-level AI risk assessment. Procurement due diligence for frontier AI products should include independent review of model cards, safety reports, red-team findings (where disclosed), and vendor incident response history. CSA's AICM Auditing Guidelines for Model Providers offer a structured framework for structuring these assessments [7].

Organizations in sectors where AI regulation is actively developing — financial services, healthcare, critical infrastructure — should track CAISI's methodology outputs. Enterprise teams should also track NIST AI RMF [8] updates, as the benchmarks and evaluation frameworks that the TRAINS Taskforce co-develops are expected to appear in future NIST guidance documents and may form the basis of sector-specific regulatory technical standards.

Strategic Considerations

The CAISI program represents one node in an emerging multi-jurisdictional AI oversight ecosystem. The UK AI Security Institute, the EU AI Act's conformity assessment regime for high-risk AI systems, and bilateral coordination frameworks between governments are converging toward a future in which frontier AI models face structured pre-deployment evaluation as a condition of market access in major

economies. Enterprises should design their AI governance programs now to be adaptable to a more demanding evaluation landscape within the coming years, as EU AI Act implementation timelines and the trajectory of NIST guidance development suggest meaningful regulatory consolidation is underway.

AI vendor relationships for frontier model procurement are increasingly treated as strategic partnerships — a shift that elevates the importance of vendor governance culture in procurement decisions. A vendor's commitment to independent security evaluation — whether through CAISI, third-party audit, or established bug bounty programs — is one indicator of governance culture. Organizations should weight this alongside technical capability assessments when making long-term AI infrastructure decisions.

CSA Resource Alignment

CAISI's pre-deployment evaluation program connects to several frameworks that enterprise teams can apply to their own AI governance programs today. CSA's resources offer one structured approach alongside complementary frameworks from other bodies; NIST AI RMF [8] and the NIST Generative AI Profile [9] offer governance structures with direct alignment to the NIST organizational context of CAISI, and enterprises should consider both alongside the CSA materials described below.

The **AI Controls Matrix (AICM) v1.0**, including auditing guidelines for Model Providers and Application Providers, addresses the shared security responsibility model across the AI supply chain. The auditing guidelines for Model Providers in particular speak to the same evaluation questions that CAISI is operationalizing at the government level — including capability assessment, safeguard testing, and independent red-teaming [7]. Enterprise teams procuring frontier AI services should use these guidelines to structure vendor assessments that go beyond published model cards.

MAESTRO (the CSA Multi-Agent Threat Modeling framework) provides a structured approach to threat modeling AI systems, including agentic AI architectures where individual model capabilities combine with tool access to create emergent risk surfaces. CAISI's focus on covert behavior, backdoors, and capability uplift is consistent with the concerns MAESTRO addresses regarding latent agent capabilities — providing a complementary analytical lens for enterprise threat modeling of the same risk surfaces CAISI evaluates at the government level.

The **Cloud Controls Matrix (CCM)** and CSA **STAR** program provide enterprise governance structures for cloud-hosted AI services. As frontier AI becomes primarily cloud-delivered, CCM controls for supply chain transparency, third-party assessment, and incident management are relevant to managing the risks that CAISI's government-facing program is designed to detect at the national security level.

CSA's ongoing work on **AI Organizational Responsibilities** addresses the governance structures that enterprises need to assign accountability for AI risk – a prerequisite for operationalizing any evaluation or compliance program, whether based on CAISI findings, NIST AI RMF [8], or sector-specific regulation.

References

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

[2] TechPolicy.Press. "Renaming the US AI Safety Institute Is About Priorities, Not Semantics." TechPolicy.Press, 2025.

[3] NIST. "U.S. AI Safety Institute Establishes New U.S. Government Taskforce to Collaborate on Research and Testing of AI Models to Manage National Security Capabilities and Risks." NIST, November 2024.

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

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

[6] Microsoft. "Advancing AI Evaluation with the Center for AI Standards and Innovation and the AI Security Institute." Microsoft On the Issues, May 2026.

[7] Cloud Security Alliance. "AICM Implementation and Auditing Guidelines." CSA AI Working Group. Accessed May 2026.

[8] NIST. "AI Risk Management Framework." NIST, January 2023.

[9] NIST. "Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile (NIST AI 600-1)." NIST, July 2024.

[10] Cybersecurity Dive. "NIST Will Test Three Major Tech Firms' Frontier AI Models for Cybersecurity Risks." Cybersecurity Dive, May 2026.