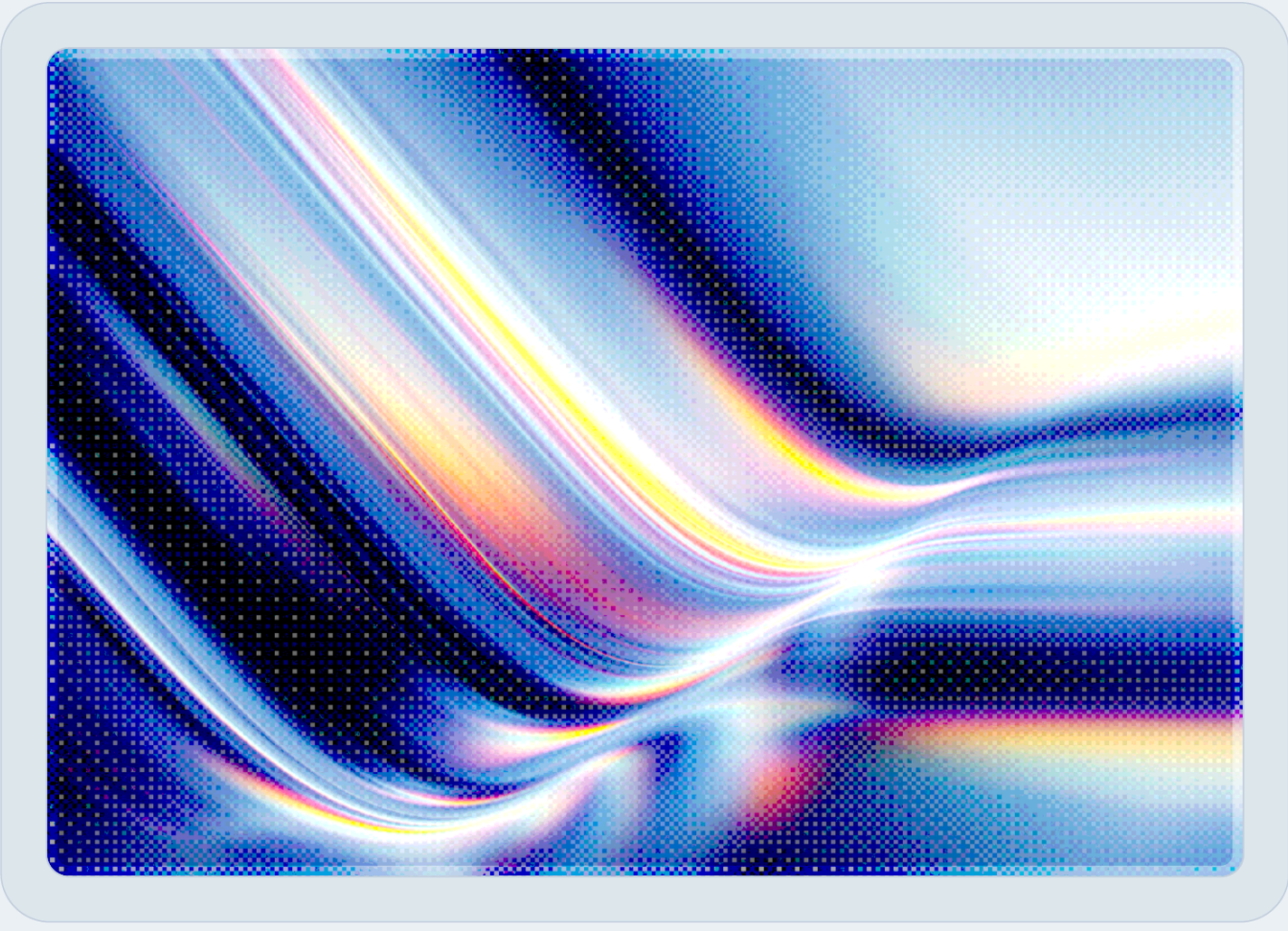
CSAI Foundation | Cloud Security Alliance

CAISI Frontier Testing Agreements Reach Five Labs

Google DeepMind, Microsoft, and xAI Join Pre-Deployment Security Review Program

2026-05-05

AI-assisted Rapid Research



CSAI | cloud security alliance®

Key Takeaways

On May 5, 2026, Bloomberg reported that Google (DeepMind), Microsoft, and xAI signed agreements with the US Center for AI Standards and Innovation (CAISI) to provide the government early access to their frontier AI models before public deployment [1]. The agreements bring the total number of frontier labs participating in CAISI's pre-deployment review program to five, with OpenAI and Anthropic having established partnerships in September 2025 [2].

- CAISI, which operates within NIST at the Department of Commerce and serves as the US government's central civilian hub for frontier AI model testing, has already completed more than 40 evaluations — including assessments of frontier models not yet available to the public [1].
- Participating developers provide CAISI with model access that includes versions with safety guardrails stripped back, enabling national security risk probing that surface-level testing would not capture [1].
- Prior evaluations conducted under existing agreements with OpenAI and Anthropic produced concrete security findings: CAISI identified novel vulnerabilities in ChatGPT Agent enabling session-scoped remote control and user impersonation, and uncovered prompt injection flaws, cipher-based evasion techniques, universal jailbreaks, and automated attack optimization paths in Anthropic's Constitutional Classifiers [2][3][4].
- The expanded roster introduces structural tensions that the current voluntary framework does not resolve. xAI has a documented history of inconsistent safety practice, while Google faces criticism from both UK lawmakers and its own workforce over perceived AI safety commitment gaps — raising questions about whether the voluntary agreement structure can produce oversight with sufficient depth and independence to match the security stakes of the models being evaluated [5][6][7].
- These agreements operate in parallel with separate Pentagon AI deals that Google, Microsoft, and xAI have each signed, a configuration that requires CAISI evaluators to navigate both security research and national security access obligations simultaneously [8][16][17].

Background

CAISI was established as the renamed successor to the US AI Safety Institute (AISI), which was itself created under the Biden administration's October 2023 executive order on AI. The Trump administration repositioned the center as CAISI, shifting emphasis toward national security and cybersecurity risk reduction. Critics characterized the rename as a substantive narrowing of scope; the administration maintained that the core evaluation mission was preserved [9]. Under the July 2025 AI Action Plan, CAISI received seventeen specific taskings spanning AI security research, national security model evaluations, global AI competition analysis, measurement science, and the development of voluntary standards [10]. The center is directed to collaborate with frontier AI developers on evaluations focused on capability risks in cyber operations, biosecurity, chemical and biological threat domains, and the integrity risks posed by adversary-developed AI systems used in critical infrastructure.

CAISI's authority to conduct these evaluations rests entirely on voluntary participation. Unlike regulatory bodies with compulsory examination powers, the center can only assess models that developers choose to share. This structure shaped how the first set of agreements was built: OpenAI and Anthropic approached CAISI after the center publicly sought AI security partners in late 2025, and the resulting partnerships were jointly scoped to focus on the model capability domains most relevant to US national security priorities [2]. In March 2026, CAISI also formalized a memorandum of understanding with the General Services Administration to extend its evaluation methodology to federal AI procurement through the USAi secure generative AI platform, positioning the center to extend its reach beyond pre-deployment assessment into evaluation of AI systems deployed in federal procurement contexts [11].

The addition of Google (DeepMind), Microsoft, and xAI on May 5, 2026 is a significant expansion in scope. These three labs — including two with AI deployments at global hyperscaler scale — collectively bring under the CAISI umbrella the underlying models that power Google Workspace, Microsoft 365 Copilot, and the Grok family of assistants, systems with user bases and enterprise integration depth that make any serious security vulnerability potentially consequential across hundreds of millions of enterprise and consumer users. OpenAI and Anthropic have simultaneously renegotiated their existing agreements to better align with AI Action Plan priorities, suggesting the program is evolving toward a more standardized framework rather than remaining a collection of ad hoc bilateral arrangements [1].

Security Analysis

What CAISI's Prior Evaluations Have Found

The September 2025 CAISI-AISI joint evaluations with OpenAI and Anthropic established a template for the type of findings the program is capable of producing when given meaningful model access. In the case of OpenAI, CAISI researchers identified two distinct novel vulnerabilities in ChatGPT Agent that, when exploited under certain conditions, enabled session-scoped remote control of the agent and user impersonation. The attack methodology chained traditional web application flaws with AI agent-specific hijacking techniques, achieving proof-of-concept exploitation at rates that demonstrated the vulnerabilities were not merely theoretical. OpenAI deployed mitigations within one business day of the findings being communicated — a notably fast turnaround that, at minimum, demonstrates the reporting channel was functional [4].

For Anthropic, CAISI and the UK AI Security Institute jointly red-teamed multiple iterations of Constitutional Classifiers, the safety filter layer applied to Claude Opus 4 and 4.1. The evaluation uncovered prompt injection attacks capable of bypassing the classifiers, cipher and obfuscation-based evasion techniques that concealed malicious intent from the filter, universal jailbreaks applicable across the model family, and automated methods for optimizing attacks against the classifier architecture. These findings informed classifier refinements before the versions in question were deployed publicly [3]. The pattern in both cases — real pre-release vulnerabilities identified and remediated before broad availability — is meaningful evidence that CAISI's access model, when fully exercised, can produce documented security improvements before broad public availability.

Risk Domains Targeted Under the New Agreements

CAISI's mandate under the AI Action Plan directs the center toward evaluation of frontier AI capabilities in a defined set of high-consequence domains. Cyber capability evaluation assesses whether models can provide meaningful uplift to adversaries conducting offensive cyber operations, including writing functional malware, identifying exploitable vulnerabilities in production software, or reasoning through network intrusion chains at a level of sophistication that exceeds what is available through non-AI means. Biosecurity evaluation tests whether models can assist in pathogen design, synthesis, or weaponization in ways that would lower the barrier to mass-casualty biological events. Chemical and nuclear domain evaluations follow similar logic. The center also conducts evaluations aimed at understanding manipulation capabilities — whether models can execute influence operations, generate persuasive disinformation at scale, or assist in systematic deception campaigns [10].

Google DeepMind's Frontier Safety Framework, now in its third iteration following an update in April 2026, maps directly onto this risk taxonomy. The framework defines Critical Capability Levels across cyber capabilities, autonomous AI research acceleration, manipulation, and chemical, biological, radiological, and nuclear threat domains, along with newly codified Tracked Capability Levels that serve as earlier warning indicators before a model approaches CCL thresholds [12]. The FSF establishes that when any model reaches a CCL, escalating security mitigations and deployment restrictions apply before external release — a policy that aligns with what CAISI evaluations are designed to verify. Whether CAISI's access under the new agreement will include pre-CCL assessment, deep red-teaming against the CCL domains, or something closer to the more limited access DeepMind previously provided to external evaluators remains to be seen.

The Access-Quality Problem

A persistent tension in frontier AI evaluation is that the depth of access determines the meaningfulness of results. A significant concern — documented in AI Lab Watch's analysis of external evaluation practices across the industry (AI Lab Watch is an advocacy organization tracking AI lab safety practices) — is that labs have in some documented cases provided external evaluators, including government safety institutes, with safety-fine-tuned model versions rather than base models, and without the ability to fine-tune or systematically probe the model at the level required to surface dangerous capabilities that safety training may mask rather than eliminate [5]. The CAISI agreement framework addresses this directly by allowing developers to hand over versions with guardrails stripped back. However, the specifics of what stripped-down access means for each lab will likely vary, and the effectiveness of the evaluation program as a whole depends on how consistently meaningful access is provided across all five participants.

xAI presents the most pointed version of this concern. The company published its initial Risk Management Framework in February 2025, and an updated final version in August 2025, both addressing malicious use and loss of control risks. According to AI Lab Watch, an advocacy organization tracking AI lab safety practices, xAI deployed a model that appeared to contradict key provisions of the August 2025 final RMF in the same week that version was published [6]. While the May 2026 CAISI agreement is a new and distinct commitment, the precedent raises legitimate questions about whether xAI's participation will yield evaluations that reflect the model capabilities actually reaching users, given the documented gap between stated safety commitments and deployed model behavior identified in its August 2025 RMF [6].

Microsoft's situation is different in character. The company has an established Deployment Safety Board, operated jointly with OpenAI, that is designed to review models before release when capability thresholds are crossed. Microsoft has also participated in the Frontier Model Forum — the industry body

it cofounded with Google, OpenAI, and Anthropic — as a venue for sharing pre-deployment risk assessment practices across labs [13]. However, Microsoft has not, to our knowledge, previously disclosed specific Deployment Safety Board review outcomes or the thresholds that triggered reviews, meaning CAISI will be working from a less established baseline than it had with Anthropic and OpenAI.

The Pentagon Overlap

A structural complexity in the current landscape is that Google, Microsoft, and xAI have all separately signed agreements with the US Department of Defense granting the Pentagon access to their models for use across lawful governmental purposes [8][16][17]. In the Google agreement, reporting has confirmed a provision enabling the government to adjust safety filter settings as needed [8]; whether comparable terms appear in the Microsoft and xAI agreements has not been publicly confirmed. Anthropic declined to sign a comparable Pentagon agreement; the DoD subsequently designated Anthropic a supply chain risk [18], though a federal court later temporarily blocked that designation. The contrasting position of the three new CAISI signatories — simultaneously agreeing to government security review of their pre-deployment models and to classified military deployment of those same models — means CAISI evaluators operate in an environment where the models they assess may be deployed in national security contexts with altered safety configurations that the evaluations themselves did not assess.

This is not a reason to dismiss the CAISI agreements as inconsequential, but it does establish that pre-deployment evaluation at CAISI and post-deployment use in classified military contexts are operating under different constraint sets. The integrity of the evaluation program depends in part on whether the model versions CAISI tests represent what will actually be deployed, or whether the safety-adjusted versions deployed in military contexts constitute a different model configuration that inherits the pre-deployment approval without having been independently evaluated.

Recommendations

Immediate Actions

Enterprise security teams building on or deploying AI systems from the five labs now under CAISI agreement should track evaluation outcomes as they become available. CAISI has published evaluation results for DeepSeek V4 Pro and has indicated it publishes findings from its assessments; forthcoming evaluations of Google, Microsoft, and xAI models may surface security characteristics relevant to

enterprise deployment decisions [14]. Organizations with existing deployments of Gemini, Copilot, or Grok should request from their vendor contacts any available documentation of CAISI evaluation scope and findings for the specific model versions they are running.

Security teams should also update their AI system procurement and third-party risk frameworks to treat CAISI evaluation status as one input among several in due diligence, rather than as a binary certification of safety. Completion of CAISI evaluation is meaningful evidence that a model has undergone structured government security review, but it is not equivalent to a comprehensive safety certification given the voluntary, scope-limited, and access-variable nature of the evaluations.

Short-Term Mitigations

Organizations deploying frontier AI systems from any of the five labs should implement runtime monitoring and behavioral controls that operate independently of model-side safety filters. The CAISI evaluations of OpenAI's systems demonstrated that safety guardrails can be bypassed at proof-of-concept rates — findings that OpenAI addressed with emergency mitigations before broader exploitation could occur. Runtime logging of agent actions, strict tool-call permission scoping, and human-in-the-loop controls for high-consequence operations remain essential security controls regardless of the pre-deployment evaluation status of the underlying model.

For organizations running AI agent deployments specifically, the session-scoped remote control and user impersonation vulnerabilities CAISI found in ChatGPT Agent underscore the importance of treating AI agent sessions as privileged execution contexts requiring the same access controls applied to privileged human users. Token isolation, per-session scope limitation, and prompt injection monitoring should be baseline requirements for any production agent deployment, regardless of the upstream model's CAISI evaluation status.

Strategic Considerations

The expansion of CAISI's evaluation roster to five major frontier labs represents a meaningful step toward institutionalizing government visibility into model capabilities before deployment, but the program's long-term security value will depend on resolution of several open architectural questions. Whether the voluntary framework can be sustained as more labs enter the market — particularly non-US developers who have no structural incentive to submit to CAISI review — will determine whether the program functions as a genuine governance mechanism or primarily as a reputational signal for companies with existing US government relationships. The current DeepSeek evaluation CAISI published

in May 2026 assesses a Chinese model's capabilities against US frontier benchmarks, but CAISI conducted that evaluation without the cooperation of the developer; that mode of evaluation does not produce the same depth of security assessment as cooperative pre-deployment access [14].

From a strategic governance perspective, the current model — voluntary agreements, unclassified evaluations, no mandatory disclosure of findings — is a reasonable starting point given political constraints, but it leaves material gaps. International alignment on pre-deployment evaluation requirements, treaty or standards-based approaches to mutual recognition between CAISI and its counterparts at the UK AI Security Institute and other allied government bodies, and legislative authority for mandatory evaluation of the most capable systems would all improve the program's coverage and enforceability. The parallel AI Seoul Summit Frontier AI Safety Commitments, to which twenty companies — including all five current CAISI partners — have subscribed, create some reputational accountability for compliance with stated safety practices, but the commitments lack audit or verification mechanisms that give external stakeholders confidence in what the labs are actually doing relative to what they have pledged [15].

CSA Resource Alignment

The developments described in this note connect to several active CSA frameworks and research workstreams. The AI Controls Matrix (AICM) provides a structured control baseline for organizations deploying AI systems that directly addresses the security evaluation and risk management practices CAISI embodies at the government level; specifically, the AICM's governance and risk management controls for AI systems are conceptually aligned with the evaluation and risk management approach CAISI applies at the government level, and organizations can use the AICM as an internal governance analogue to assess whether their own AI procurement and deployment practices achieve comparable evaluation depth. The AICM is the recommended primary framework for AI governance risk and compliance, superseding standalone Cloud Controls Matrix application for AI-specific contexts.

The CSA MAESTRO framework for agentic AI threat modeling is directly relevant to the ChatGPT Agent vulnerabilities CAISI identified — prompt injection, tool hijacking, and session-scoped credential abuse are precisely the threat categories MAESTRO's threat modeling methodology is designed to surface in enterprise agentic deployments. Organizations that have not conducted MAESTRO-aligned threat modeling for their AI agent deployments should treat the CAISI findings from the OpenAI evaluation as empirical confirmation that the threat categories MAESTRO addresses are not theoretical.

CSA's AI Organizational Responsibilities publications — specifically the volumes addressing Governance, Risk Management, Compliance, and Cultural Aspects — address the governance framework design questions that the CAISI voluntary agreement model raises at scale. The Responsibility framework's guidance on AI vendor assessment, third-party AI risk management, and board-level AI governance reporting provides a reference model for how enterprise organizations should translate CAISI evaluation results and limitations into their own internal risk posture.

The STAR (Security Trust Assurance and Risk) program provides a structured assurance mechanism that organizations can use to evaluate AI providers' own security practices, complementing the capability-focused evaluation CAISI conducts. Frontier AI labs that publish STAR-level assessments of their AI development and deployment security practices give enterprise customers a more complete picture of both the capability risks CAISI is assessing and the operational security practices surrounding those capabilities.

References

[1] Bloomberg. "AI Firms to Give US Government Early Access for Model Evaluation." Bloomberg Technology, May 5, 2026.

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

[3] Anthropic. "Strengthening Our Safeguards Through Collaboration with US CAISI and UK AISI." Anthropic, September 2025.

[4] OpenAI. "Working with US CAISI and UK AISI to Build More Secure AI Systems." OpenAI, October 2025.

[5] AI Lab Watch. "AI Companies Aren't Really Using External Evaluators." AI Lab Watch, May 2024.

[6] AI Lab Watch. "xAI's New Safety Framework Is Dreadful." AI Lab Watch Substack, September 2025.

[7] Time Magazine. "Exclusive: 60 U.K. Lawmakers Accuse Google of Breaking AI Safety Pledge." Time, 2025.

[8] The Hill. "Google, Pentagon Strike Deal for Artificial Intelligence Services." The Hill, April 2026.

[9] SSTI. "The U.S. AI Safety Institute Has Been Renamed the Center for AI Standards and Innovation." SSTI, 2025.

[10] White House. "America's AI Action Plan." White House, July 2025.

[11] NIST/CAISI. "CAISI Signs MOU with GSA to Boost AI Evaluation Science in Federal Procurement Through USAi." NIST, March 2026.

[12] Google DeepMind. "Strengthening Our Frontier Safety Framework." Google DeepMind, September 2025 (updated April 2026).

[13] Frontier Model Forum. "About the Frontier Model Forum." Frontier Model Forum, 2023.

[14] NIST/CAISI. "CAISI Evaluation of DeepSeek V4 Pro." NIST, May 2026.

[15] UK Government. "Frontier AI Safety Commitments, AI Seoul Summit 2024." GOV.UK, updated February 2025.

[16] TechCrunch. "Microsoft Signs AI Agreement with the Pentagon." TechCrunch, May 2026.

[17] Axios. "xAI Signs AI Agreement with the Department of Defense." Axios, February 2026.

[18] TechCrunch. "It's Official: The Pentagon Has Labeled Anthropic a Supply-Chain Risk." TechCrunch, March 2026.