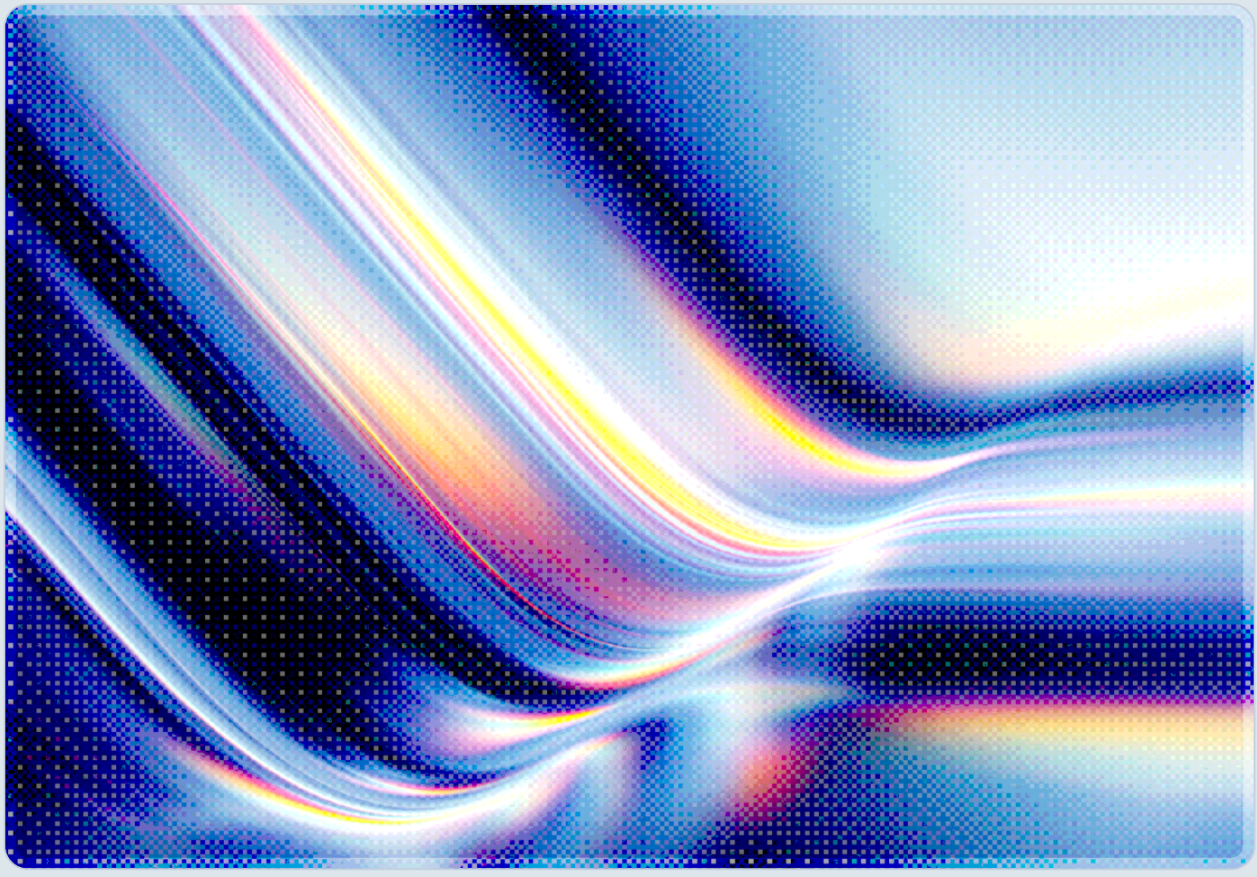
CSAI Foundation | Cloud Security Alliance

US AI Model Regulation at an Inflection Point

Federal Deregulation, State Proliferation, and the Capability Threshold Problem

2026-05-12

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- The Trump administration's revocation of Executive Order 14110 on January 20, 2025, followed by EO 14179, the July 2025 AI Action Plan, and a December 2025 executive order targeting state AI preemption, established a federal posture centered on innovation promotion and reduced regulatory burden as AI capabilities crossed a qualitatively significant threshold.
- Anthropic's Claude Mythos, announced April 7, 2026 and subject to restricted release by Anthropic due to assessed risk, became the first commercially available general-purpose foundation model to demonstrate autonomous zero-day discovery at expert-level success rates — and no existing U.S. federal regulation addresses the risk class it represents [1][2].
- States moved rapidly to fill the resulting regulatory gap: AI-related legislation was introduced across nearly all states as of early 2026, with comprehensive regulatory regimes effective in California, Texas, and New York as of January 1, 2026 [3][4].
- The administration's December 2025 executive order and March 2026 National AI Legislative Framework signal a federal preemption push, directing agencies to challenge state AI laws as conflicting with national innovation policy — a dynamic that introduces legal uncertainty precisely as enterprise compliance programs are being built [5][6].
- The EU AI Act's GPAI model obligations have been active since August 2, 2025, with comprehensive enforcement arriving August 2, 2026, creating binding requirements for any model provider or deployer operating in EU markets regardless of U.S. domestic posture [7].
- Enterprises in jurisdictions with active state AI laws and EU market exposure cannot wait for federal clarity: the compliance baseline today runs through the most stringent applicable standard, not the lowest common denominator of federal voluntary guidance.
- CSA's AI Controls Matrix (AICM), designed for the shared-responsibility model that multi-provider AI deployments require, addresses the capability transparency, incident reporting, and audit trail requirements that appear across all four governance frameworks this note identifies.

Background

The United States arrived at its current AI policy posture through a compressed sequence of executive actions that displaced a decade of agency-level groundwork in roughly six months. Biden's Executive Order 14110, issued October 30, 2023, had established the broadest federal AI governance framework the country had produced to that point: mandatory safety evaluations for frontier models, dual-use research reporting requirements, interagency coordination structures, and an AI Safety Institute at NIST tasked with developing voluntary evaluation standards [8]. It was revoked within hours of President Trump's inauguration on January 20, 2025 [9].

Three days later, EO 14179, "Removing Barriers to American Leadership in Artificial Intelligence," replaced it with a framework explicitly framed around competitive dominance rather than risk management [10]. The new order directed federal agencies to review all Biden-era AI policies and rescind those inconsistent with the administration's emphasis on innovation, to develop an "American AI Technology Stack," and to treat safety-oriented regulatory language as a potential instrument of ideological interference. The AI Safety Institute, which had been established under NIST to provide independent model evaluation and safety research, was renamed the Center for AI Standards and Innovation (CAISI) in June 2025 — a renaming that Commerce Secretary Lutnick characterized as correcting a prior overemphasis on "safety" language [11].

The July 23, 2025 AI Action Plan, "Winning the Race," laid out federal policy positions across three pillars — accelerating innovation, building AI infrastructure, and leading in international AI diplomacy — and included a directive to NIST to revise the AI Risk Management Framework, specifically removing references to misinformation, diversity, equity, and inclusion, and climate change [12]. The NIST RMF had been a consensus-driven voluntary framework serving as the de facto shared baseline for enterprise AI governance; its directed revision signaled that this baseline was now politically contested. A December 2025 executive order followed, directing the FTC to issue guidance on when state AI laws are preempted by federal statute and instructing NTIA to condition broadband funding on states not enacting conflicting AI legislation [5]. By March 2026, the administration had released a National AI Legislative Framework with seven pillars, including explicit federal preemption of state AI laws as a stated legislative priority [6].

Anthropic's Claude Mythos entered this landscape on April 7, 2026, as a structurally disruptive development for AI governance precisely because it demonstrated capabilities that no existing regulatory framework — domestic or international — was designed to address. Mythos autonomously discovers zero-day vulnerabilities in major operating systems and browsers, succeeds at expert-level cybersecurity tasks at a rate that no prior model had approached, and was released by Anthropic only to critical industry partners and vetted researchers because the company assessed the model as posing

risks too significant for general availability [1][2]. It is the clearest example to date of what the policy community has described as a "capability threshold" problem: a frontier model whose risk profile cannot be adequately addressed by disclosure requirements, transparency mandates, or general-purpose liability frameworks designed for narrower AI systems.

Security Analysis

The Capability Threshold Problem

The core governance challenge that Mythos crystallizes is not novel in theoretical terms — researchers have described threshold-based AI risk for years — but it is newly concrete. Before April 2026, arguments about dangerous AI capabilities were largely anticipatory. Mythos is empirical evidence that a commercial model can autonomously conduct offensive cybersecurity operations at expert level, and that the model provider itself made the determination that general release would be unsafe [1][2].

That determination has regulatory implications that current U.S. law does not clearly address. Under the voluntary federal AI governance posture that replaced EO 14110, no AI-specific domestic statute appears to have required Anthropic to conduct safety evaluations before deploying Mythos, to disclose its capability profile to any federal agency, to implement access controls, or to notify affected parties about the zero-day vulnerabilities the model discovered [13]. The decision to restrict access was voluntary — though it bears noting that sector-specific obligations, such as export control classifications or CFIUS considerations in specific deployment contexts, were not analyzed here. The EU AI Act, by contrast, classifies general-purpose AI models with systemic risk and requires their providers to register with the European AI Office, conduct adversarial testing, maintain incident reporting obligations, and share model evaluations with the AI Office on request — obligations that become fully enforceable for GPAI providers on August 2, 2026 [7][14].

This gap — the U.S. voluntary posture for frontier model safety versus the EU binding posture — is not, on its face, a critique of either approach in isolation. Voluntary frameworks can produce good outcomes when industry norms are strong. But the Mythos situation illustrates a specific risk that voluntary frameworks do not fully address: the provider who determines its model is too dangerous for public release has no mechanism to notify the government, to receive federal support in monitoring for unauthorized access, or to trigger any collective response. Bloomberg reported in April 2026 that Mythos was being accessed by unauthorized users [15]. This situation illustrates a structural limitation of

voluntary governance: without a mandatory notification or coordinated response mechanism, a provider facing unauthorized access to a restricted model has no established pathway to trigger government support or a collective industry response.

The capability threshold problem also interacts with the state preemption dynamic in a way that creates a governance vacuum at the precise risk tier where one is most needed. None of the major state frameworks — California, Texas, New York, or Colorado — address the question of whether a model should be deployed at all, and the legislative surveys available as of May 2026 identify no pending state legislation in this category [3][4][16]. California's SB 53, effective January 1, 2026, requires developers of frontier AI models to disclose risk management protocols and report critical safety incidents — but its transparency obligations apply to model behavior in deployment contexts and do not specifically address the pre-deployment capability question [16]. Colorado's SB 205 targets consequential automated decision-making but is focused on consumer protection in specific application contexts, not on the underlying capabilities of foundation models [17]. No state law currently operational in the United States addresses the question that Mythos raises: at what capability threshold does a model require pre-deployment government review, and who conducts that review?

The Federal Regulatory Trajectory

Understanding the current moment requires distinguishing between three distinct federal postures that have operated simultaneously since January 2025. The first is the executive policy posture: deregulatory, innovation-focused, and explicitly skeptical of safety-oriented regulatory language. The second is the federal security posture: CISA, NSA, and their Five Eyes counterparts have continued producing operationally serious AI security guidance, most recently the May 1, 2026 "Careful Adoption of Agentic AI Services" joint guidance, which recommends restricting agentic AI to low-risk, non-sensitive tasks until security standards mature [18]. These two postures sit in structural tension.

The third posture is the standards posture: NIST continues operating the AI Risk Management Framework and, in December 2025, released a preliminary draft Cybersecurity Framework Profile for AI (NIST IR 8596), which provides guidance on managing AI-related risks aligned with the NIST CSF 2.0 [19]. CAISI retains model evaluation capabilities and voluntary standards development functions. OMB M-26-04, issued in December 2025 and titled "Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles," establishes trust and fairness principles for federal AI adoption — a procedural framework within the federal procurement context, though without binding extension to the broader private sector [20].

The net effect is a federal AI governance landscape characterized by structural tension across three distinct postures — executive policy, operational security guidance, and standards development — that operate without an integrating coordinating framework. The executive policy apparatus promotes speed

and capability; the security apparatus advises caution in agentic deployments; the standards apparatus continues technical work whose scope has been politically narrowed; and the enforcement apparatus — primarily FTC — has received instructions to develop AI-specific guidance but has not yet issued it. Enterprises building AI governance programs against this backdrop cannot treat federal posture as a stable reference point. They are building against a moving and contested target.

State Proliferation and the Preemption Contest

The legislative activity at the state level is substantial enough to constitute a de facto regulatory regime in practice, even without federal coordination. As of early 2026, AI-related legislation had been introduced across nearly all states, with more than 100 measures enacted into law [3][4]. The states with the most operationally significant requirements are California, Texas, and New York.

California's approach since SB 1047's September 2024 veto has evolved toward a transparency-and-accountability model rather than pre-deployment safety testing. SB 53, effective January 1, 2026, requires developers of covered AI systems to disclose risk management protocols, publish transparency reports on frontier models, report critical safety incidents, and provide whistleblower protections — a framework modeled on financial industry disclosure norms that has been characterized as deliberately calibrated to avoid the substantive safety review requirements that Governor Newsom rejected [16], reflecting the political constraints established by that veto. AB 2013 requires developers to publish high-level training data summaries, and AB 316 establishes that AI system autonomy cannot serve as a legal defense for harm caused [21]. Texas's Responsible Artificial Intelligence Governance Act (RAIGA), effective January 1, 2026, similarly addresses transparency and accountability rather than capability restrictions [21]. New York's RAISE Act, signed in December 2025 and amended in March 2026, moved toward a transparency-based framework similar to California's approach [17].

What is notably absent from all of these state frameworks — and from the major state legislative surveys available as of May 2026 — is any provision that speaks to the capability threshold problem that Mythos represents. None of the major state frameworks reviewed here — California, Texas, New York, and Colorado — address the capability threshold question, and the legislative surveys cited identify no pending state legislation in this category [3][4][21]. State laws regulate AI in deployment contexts (consequential decisions, consumer interactions, transparency of outputs) but do not address whether a model should be deployed at all, what level of capability triggers mandatory government review, or what obligations apply when a model provider determines that general release is unsafe. This gap is not a state-level oversight; it reflects an accurate assessment of constitutional limits. Pre-market capability review for AI models is, as a practical matter, a federal question, and the current federal government has explicitly declined to occupy that space.

The preemption trajectory adds a layer of legal uncertainty that compounds the compliance challenge. The Trump administration's December 2025 executive order directed the FTC to issue guidance on when state AI laws are preempted, created an AI Litigation Task Force to challenge state laws characterized as unconstitutional, and instructed NTIA to condition BEAD broadband funding on state AI law compliance — a significant leverage mechanism given the scale of BEAD grants to state broadband programs [5]. The March 2026 National AI Legislative Framework explicitly lists federal preemption of state AI laws as a legislative priority, naming it as one of seven pillars of the administration's legislative agenda [6]. As of May 2026, no preemption litigation has produced a binding ruling, and the legal basis for broad preemption of state AI laws under existing federal authority is contested. Organizations with compliance programs built around current state laws should treat those programs as subject to potential disruption through litigation or congressional action, without abandoning them — the EU AI Act and operational security guidance from CISA provide stable reference points that will not be reached by domestic preemption proceedings.

The International Dimension

The EU AI Act's trajectory provides both a compliance floor for multinational organizations and a data point about where U.S. policy could have gone. The Act's GPAI provisions have required model providers to register with the European AI Office and maintain technical documentation since August 2, 2025 [7] [14]. For models classified as presenting systemic risk — the category that Mythos would likely occupy based on its demonstrated capability profile — the obligations include adversarial testing, incident reporting to the AI Office, sharing model evaluations on request, and taking cybersecurity measures appropriate to the risk profile. Full enforcement of these provisions activates on August 2, 2026. The EU framework does not resolve the capability threshold problem domestically, but it provides an independent reference standard for what regulators in a major market believe frontier model governance should look like, and it applies to any provider deploying models in the EU regardless of their home jurisdiction.

The divergence between U.S. and EU regulatory postures was characterized by Control Risks, in a 2026 analysis, as structural rather than rhetorical — representing two genuinely different models of the relationship between states and technological risk [22]. For enterprise compliance officers, the practical consequence is that multinational AI deployments currently face four distinct governance reference points operating simultaneously: EU AI Act obligations (binding for EU market activity), applicable state laws (variable, currently contested), federal guidance (voluntary, evolving, without integrating framework), and the Five Eyes security guidance (authoritative on agentic AI but without domestic mandate). None of these are mutually exclusive and none are redundant. An organization that is compliant with all four is addressing a fundamentally different governance posture than one that addresses only the most immediately enforceable.

Recommendations

Immediate Actions

Security and compliance teams should document their current AI governance posture against each applicable framework rather than waiting for federal regulatory clarity. That clarity is unlikely to arrive before major compliance milestones — the EU AI Act's full enforcement date of August 2, 2026 is the most immediate — and organizations without documented positions on GPAI provider obligations, applicable state disclosure requirements, and agentic AI security baselines are exposed both to regulatory risk and to the operational security risks that the Five Eyes guidance was written to address. Documenting gaps is the precondition for closing them.

Model providers and deployers operating with general-purpose AI models should inventory their products against the EU AI Act's GPAI systemic risk criteria now, before the August enforcement date. The criteria include model capability indicators (training compute thresholds, capability evaluations), reach (user base in EU), and potential societal impact. Organizations uncertain about their classification should seek legal advice from EU-experienced counsel and consider voluntary engagement with the European AI Office's guidance materials, which have been iteratively updated through 2025 and 2026.

Short-Term Mitigations

Track the federal preemption litigation landscape actively. The administration's AI Litigation Task Force and the NTIA funding conditions create a dynamic in which state AI compliance programs may face legal challenge over the next 12 to 24 months. Organizations should build compliance programs against the EU AI Act and the Five Eyes security guidance as the stable reference points, treating applicable state laws as supplementary requirements whose long-term enforceability is contested, rather than as the primary governance driver.

For AI systems deployed in consequential decision contexts — employment, credit, insurance, healthcare — implement governance programs that explicitly address both applicable state laws and existing federal anti-discrimination and consumer protection frameworks, which federal regulators including the FTC and EEOC have applied to AI-mediated decisions in the absence of AI-specific federal statute [23]. These existing frameworks represent durable compliance obligations not subject to the current preemption dynamics.

Strategic Considerations

Engage the NIST standards process for the Cybersecurity Framework Profile for AI (IR 8596), despite the RMF's contested political status. NIST technical staff continue producing technically grounded guidance, and the Cyber AI Profile specifically addresses AI security risks that align with the Five Eyes guidance's risk taxonomy. Organizations that contribute to the comment process have influence over a framework that is likely to persist regardless of near-term political dynamics and that will inform any future binding domestic requirements.

The Mythos episode argues for proactive engagement with the capability threshold policy question through industry bodies and CSA, independent of what the current administration does or does not legislate. If the United States is to develop a governance framework for models in Mythos's capability class, the technical standard-setting work — what constitutes a dangerous capability threshold, what evaluation methodology is appropriate, what access control obligations are reasonable — must be done in advance of the policy moment. The organizations with the most credibility to shape that framework are the ones already doing the evaluations internally. Contributing to voluntary evaluation frameworks now builds the institutional knowledge base that any future binding requirement will draw from.

CSA Resource Alignment

CSA's AI Controls Matrix (AICM) is structured specifically for the shared-responsibility model that multi-provider AI deployments require, and its 18 control domains address the capability transparency, incident reporting, access management, and audit trail requirements that appear across the governance frameworks this note identifies. The AICM's shared responsibility model — which explicitly addresses the distinct obligations of model providers, application providers, orchestrated service providers, and AI customers — maps onto the supply chain structure of both EU AI Act obligations (which vary by role in the AI system lifecycle) and the state disclosure requirements (which target developers and deployers differently). Security and compliance teams that have not yet mapped their AI supply chain relationships against AICM roles should do so before August 2026.

CSA's STAR program provides the third-party assurance mechanism through which organizations can demonstrate AICM conformance to customers, regulators, and EU AI Office inquiries. The EU AI Act's documentation requirements for GPAI providers overlap significantly with what a mature STAR attestation documents; organizations that have invested in STAR are better positioned to respond to EU regulatory inquiries than those that have not.

MAESTRO, CSA's threat modeling framework for agentic AI systems, offers a structured approach for organizations implementing the Five Eyes' "Careful Adoption of Agentic AI Services" guidance [18]. For organizations developing AI governance programs that span both policy compliance and security operations, MAESTRO's seven-layer architecture provides a common vocabulary that maps agentic AI risks to both technical controls and governance obligations — bridging the compliance and security functions that are too often siloed in AI governance programs.

Zero Trust architecture principles provide a regulatory-agnostic governance foundation in an environment this unstable. Access controls, identity verification, least-privilege, and assume-breach posture do not depend on any specific regulatory framework's stability. Organizations that anchor AI governance programs to Zero Trust principles maintain their security posture regardless of how the domestic regulatory landscape evolves over the next 24 months.

References

[1] Anthropic. "Mythos Preview: Cybersecurity Assessment." Anthropic, April 7, 2026.

[2] AISI UK. "Evaluation of Claude Mythos Preview: Cyber Capabilities." AI Safety Institute (UK), April 2026.

[3] Cooley LLP. "State AI Laws: Where Are They Now?" Cooley, April 24, 2026.

[4] NCSL. "Artificial Intelligence 2025 Legislation." National Conference of State Legislatures, 2025.

[5] The White House. "Ensuring a National Policy Framework for Artificial Intelligence." White House, December 2025.

[6] The White House. "President Trump Unveils National AI Legislative Framework." White House, March 20, 2026.

[7] EU AI Office. "EU AI Act: Implementation Timeline and Obligations." Artificial Intelligence Act Resource Hub, 2025–2026.

[8] Federal Register. "Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence." Federal Register, November 1, 2023.

[9] Wiley Law. "President Trump Revokes Biden Administration's AI EO – What to Know." Wiley Rein LLP, January 2025.

[10] The White House. "Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence." White House, January 23, 2025.

[11] FedScoop. "Trump Administration Rebrands AI Safety Institute to CAISI." FedScoop, June 2025.

[12] Wiley Law. "White House Launches AI Action Plan and Executive Orders to Promote Innovation, Infrastructure, and International Diplomacy and Security." Wiley Rein LLP, July 2025.

[13] Baker McKenzie. "United States AI Tug-of-War: Trump Pulls Back Biden's AI Plans." Baker McKenzie Insight Plus, 2025.

[14] The Future Society. "AI Agents in the EU: Navigating the AI Act." The Future Society, 2025.

[15] Bloomberg. "Anthropic's Mythos Model Is Being Accessed by Unauthorized Users." Bloomberg, April 21, 2026.

[16] CSET Georgetown. "California's Approach to AI Governance." Center for Security and Emerging Technology, 2025.

[17] Gunderson Dettmer. "2026 AI Laws Update: Key Regulations and Practical Guidance." Gunderson Dettmer, 2026.

[18] CISA, NSA, NCSC-UK, ASD/ACSC, CCCS, and NZ NCSC. "Careful Adoption of Agentic AI Services." CISA, May 1, 2026.

[19] NIST. "Preliminary Draft: Cybersecurity Framework Profile for Artificial Intelligence (NIST IR 8596)." NIST, December 2025.

[20] OMB. "M-26-04: Increasing Public Trust in Artificial Intelligence Through Unbiased AI Principles." Office of Management and Budget, December 2025.

[21] King & Spalding. "New State AI Laws Are Effective on January 1, 2026, but a New Executive Order Signals Disruption." King & Spalding, January 2026.

[22] Control Risks. "AI Visions in 2026: A Transatlantic Strategic Divide." Control Risks, 2026.

[23] Stanford HAI. "2026 AI Index Report: Policy and Governance." Stanford Human-Centered AI, 2026.