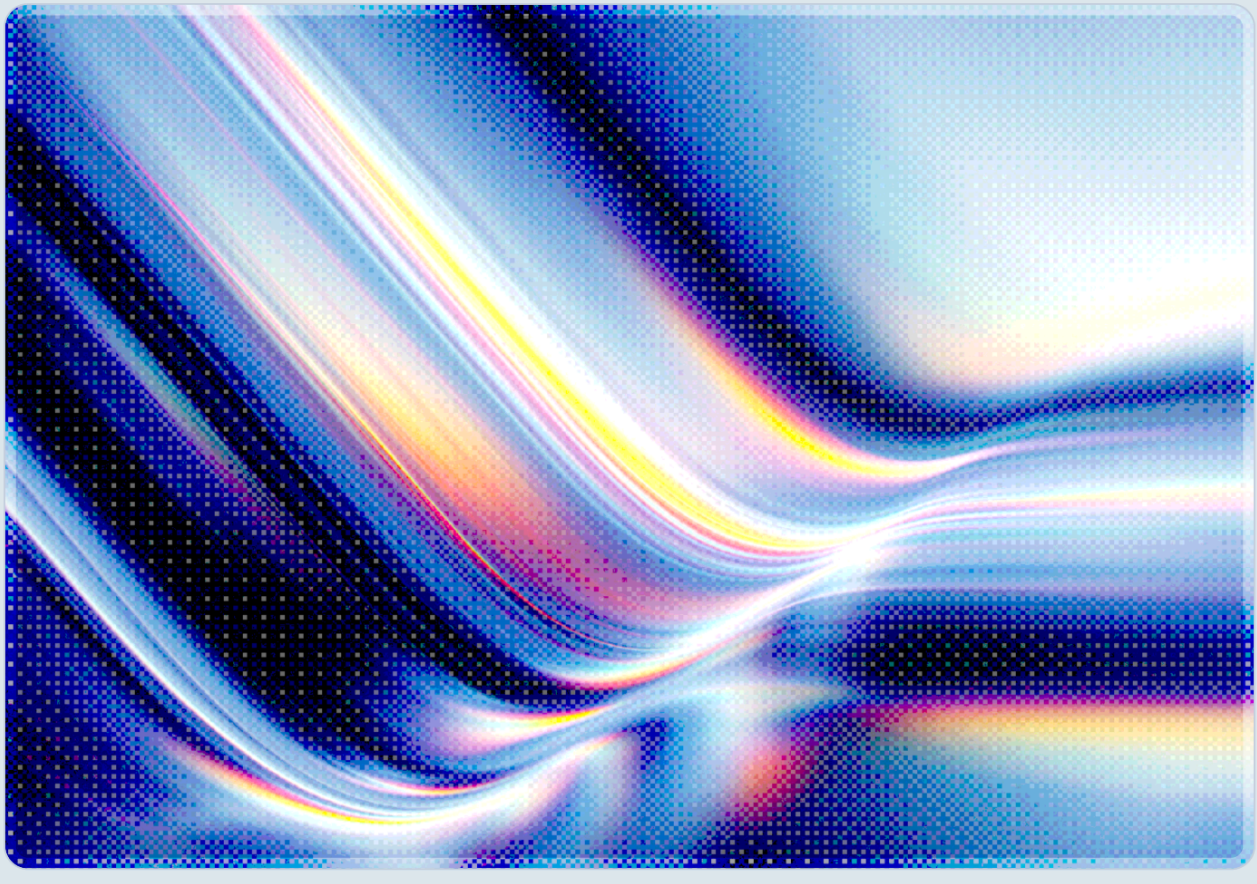
CSAI Foundation | Cloud Security Alliance

Agentic AI Governance at a Crossroads

Five Eyes Guidance Confronts U.S. Regulatory Rollback

2026-05-10

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

- On May 1, 2026, six allied cybersecurity agencies—CISA, NSA, and the national cyber centers of the United Kingdom, Australia, Canada, and New Zealand—jointly published "Careful Adoption of Agentic AI Services," the first publicly available joint guidance from six Five Eyes-affiliated agencies specifically addressing the security risks of autonomous AI agent systems [1][2].
- This guidance arrives against a consistent U.S. policy trajectory that revoked the principal federal AI safety order (EO 14110) on January 20, 2025, renamed the AI Safety Institute to the Center for AI Standards and Innovation (CAISI), and directed NIST to revise the AI Risk Management Framework to remove specific risk-oriented content [3][6][7].
- The result is a structural governance gap: U.S. intelligence and security agencies are co-authoring operationally authoritative agentic AI security guidance with allied nations, while the civilian policy apparatus that would implement such guidance domestically has been substantially reduced in scope.
- The Five Eyes guidance identifies five risk categories specific to agentic AI—privilege escalation, design and configuration flaws, behavioral unpredictability, cascading structural failures, and accountability deficits—which it characterizes as inadequately covered by existing AI governance frameworks [1].
- Security and compliance professionals should treat the joint guidance as a de facto operational standard for agentic AI deployments, particularly where no domestic mandate currently exists. CSA's MAESTRO framework and the December 2025 OWASP Top 10 for Agentic Applications provide the technical implementation layer against which this policy guidance should be evaluated.

Background

The term "agentic AI" describes AI systems that operate with meaningful autonomy: they receive high-level objectives, decompose those objectives into steps, select and invoke external tools, persist decisions across multiple reasoning cycles, and take consequential actions in the world with limited or intermittent human oversight. Unlike a static language model that responds to a single prompt and

returns a single output, an agentic system may orchestrate dozens of tool calls, spawn subordinate agents, write and execute code, query databases, send messages, and modify infrastructure—all in pursuit of a goal specified once at invocation. The architecture amplifies both the productive potential and the attack surface of AI considerably [1].

The threat model for agentic AI differs in kind, not merely degree, from traditional software security. A compromised web application executes attacker-controlled instructions; a compromised AI agent reasons about how to pursue attacker-controlled goals using all available capabilities, including capabilities its operators may not have fully inventoried. The autonomy that makes agentic systems valuable also makes them harder to audit, harder to constrain, and harder to attribute when something goes wrong. Multi-agent architectures compound these difficulties: when an orchestrating agent delegates tasks to worker agents, the decision graph expands rapidly, and a manipulation injected at one layer may propagate unpredictably across the network before any human observer can intervene.

This risk profile has reached a threshold requiring coordinated governmental response, as evidenced by the publication of "Careful Adoption of Agentic AI Services" by six Five Eyes-affiliated agencies on May 1, 2026 [1][2]. The guidance is notable for what it is as much as for what it says: among the first instances in which the allied signals intelligence community treated agentic AI as a distinct security domain requiring its own threat taxonomy and mitigation framework, it signals that agentic risk can no longer be addressed as a subset of general AI or cloud security concerns. This paper uses the term "structural governance gap" to describe the condition in which domestic civilian regulatory capacity has been reduced in scope precisely as the technical risk profile those institutions were designed to address has grown more acute and internationally prominent.

The timing of this publication places it squarely within a contested policy moment in the United States. The guidance from CISA and NSA is operationally authoritative in the sense that it reflects the expertise of the agencies responsible for federal civilian cybersecurity and national signals intelligence—but it arrives without a corresponding domestic regulatory mandate to implement it. Understanding why requires tracing the U.S. AI policy trajectory over the preceding sixteen months.

Security Analysis

The Five Eyes Guidance: Scope and Key Findings

The "Careful Adoption of Agentic AI Services" guidance organizes the risks of agentic AI into five categories, each of which reflects a characteristic that distinguishes autonomous agents from prior AI deployment patterns [1].

The first category is privilege. The guidance observes that agentic systems are commonly granted broad access rights because their tasks require interacting with diverse systems and data sources, and that this access is rarely scoped to the minimum necessary for any given task—meaning that a single agent compromise grants an attacker a wide operational footprint. The recommendation is not simply to reduce permissions but to implement dynamic, just-in-time privilege allocation that grants and revokes access at the task level rather than the system level—a materially different model from how most organizations currently manage service account permissions.

Design and configuration flaws constitute the second category. Many agentic deployments inherit security assumptions from the frameworks and APIs they consume without validating that those assumptions hold in the operational context. Configuration decisions made at deployment time—which tools an agent can invoke, which credentials it carries, which systems it can modify—may often reflect convenience rather than deliberate security design. The guidance emphasizes that because agentic systems act on behalf of organizations in consequential ways, configuration decisions warrant the same scrutiny as infrastructure-level security architecture.

The third category addresses behavioral risk: agents that pursue their stated objectives in ways that produce outcomes their operators did not anticipate or sanction. This category resists the traditional software security framing of "intended versus unintended behavior," because an agent operating exactly as designed may still take actions that surprise its operators when operating in edge conditions or novel environments. Behavioral risk also encompasses prompt injection—where malicious instructions embedded in data the agent processes cause it to take unauthorized actions—and goal generalization failures in multi-agent systems.

Structural risk, the fourth category, refers to the cascading failure modes that emerge in multi-agent architectures. When agents are interconnected, a false signal, a corrupted tool response, or a compromised subordinate agent can propagate through the network in ways that are difficult to detect and interrupt before significant damage occurs. The guidance explicitly notes that the speed of automated decision-making in these systems reduces the window for human intervention compared to traditional software pipelines.

Accountability deficits constitute the fifth category, and in some respects the most consequential for security governance. When an agentic system takes an action—deletes a file, sends a message, modifies a record, executes a transaction—the causal chain leading to that action may span multiple model calls, tool invocations, and agent-to-agent communications. The guidance notes that existing logging and auditing practices can be inadequate to reconstruct this chain. Without interpretable audit trails, incident response is impaired, regulatory obligations are difficult to demonstrate, and the lessons of failures cannot be systematically applied.

The headline recommendation of the guidance is that organizations should restrict agentic AI to "low-risk, non-sensitive tasks" until security standards and operational practices mature [1][2]. This is a conservative posture that reflects how early the field is in developing validated defenses, and it stands in deliberate tension with the commercial pressure to deploy agentic systems rapidly and broadly.

The U.S. Regulatory Context

The policy environment in which this guidance arrives is defined by a series of executive actions beginning on January 20, 2025, when President Trump revoked Executive Order 14110, the Biden administration's October 2023 comprehensive AI safety order [3][4][5]. EO 14110 had established federal requirements for safety evaluations, dual-use testing, cross-agency reporting, and standards development. Its revocation was followed three days later by Executive Order 14179, "Removing Barriers to American Leadership in Artificial Intelligence," which directed federal agencies to review all Biden-era AI policies and suspend, revise, or rescind those inconsistent with a new emphasis on innovation, competitiveness, and reduced regulatory burden [3][19].

The organizational consequences materialized over the following months. In February 2025, the director of the AI Safety Institute, Elizabeth Kelly, departed amid broader federal workforce reductions [20]. On June 4, 2025, Commerce Secretary Howard Lutnick announced the renaming of AISI to the Center for AI Standards and Innovation (CAISI), framing the prior "safety" terminology as language used, in his characterization, under the guise of security concerns to justify overreach and ideological interference [7][8]. CAISI retained some operational functions—voluntary standards development, model evaluations—but the mission reframing was significant as a signal about how the administration views the relationship between AI capability and AI risk.

The AI Action Plan released July 23, 2025, titled "Winning the Race: America's AI Action Plan," set out three pillars—accelerating AI innovation, building AI infrastructure, and leading in international AI diplomacy and security—and accompanied three executive orders addressing AI exports, data center permitting, and the content requirements of government-contracted AI systems [6]. The Action Plan's directive to NIST is particularly relevant to the governance gap identified here: NIST was instructed to revise the AI Risk Management Framework to eliminate references to misinformation, diversity, equity, and inclusion, and climate change [6]. This represents a content change to a framework developed through NIST's multi-stakeholder process and widely adopted as a voluntary governance baseline [14], and it signals that the RMF's role as a shared reference is now subject to political revision.

The effect on the private sector is not primarily through direct regulation—the U.S. retains a largely voluntary domestic AI governance framework—but through the institutional infrastructure that supports voluntary compliance. Organizations that previously used NIST guidance, AISI evaluations, and federal

interagency coordination as reference points for their AI security programs now face a reference landscape that is more fragmented, more contested, and less aligned with allied partner frameworks than it was before January 2025.

The International Divergence

The Five Eyes guidance exists within a broader international policy landscape that is increasingly bifurcated. The EU AI Act's obligations for general-purpose AI models entered enforcement on August 2, 2025; its high-risk system requirements, which explicitly apply to autonomous agents operating in high-stakes contexts, take effect August 2, 2026 [11][21]. The EU framework is legally binding within the bloc and creates compliance obligations for any organization deploying AI systems to EU markets, regardless of where those systems are developed. Analysts from Control Risks, writing in early 2026, characterized the U.S.-EU divergence as "structural" rather than rhetorical, representing two genuinely different models of how states relate to technological risk [12]. Chatham House's March 2026 analysis identified bloc formation around these two models as a primary barrier to coherent global AI governance at a moment when the technology's pace demands coordination [13].

The intermediate positions are instructive. Australia reversed course in December 2025, announcing through its National AI Plan that it would abandon proposed mandatory AI guardrails in favor of voluntary guidance and reliance on existing technology-neutral laws—a structural alignment with U.S. posture, even as the Australian Signals Directorate co-signed the Five Eyes guidance [15][18]. Canada's proposed Artificial Intelligence and Data Act (Bill C-27) died when Parliament was prorogued on January 6, 2025, terminating the bill before the INDU Committee could complete its study [17]. The United Kingdom, under the Labour government, has maintained a principles-based, non-legislative approach to AI regulation and has not backed the private member's AI regulation bill introduced in the House of Lords in March 2025 [16].

The most significant near-term consequence of this divergence is not regulatory but operational. Multinational enterprises deploying agentic AI systems simultaneously face the EU AI Act's binding requirements, the Five Eyes guidance's strong operational recommendations, and the absence of equivalent domestic mandates in the U.S., Canada, and Australia. For organizations in this position, the practical compliance path runs through the most stringent applicable standard—EU obligations—while the Five Eyes guidance provides the security-specific implementation layer the EU framework does not prescribe. The net effect is de facto convergence through market behavior rather than policy coordination, but this convergence is uneven, ungoverned, and is likely to impose higher compliance burdens than a coordinated international framework would require.

Recommendations

Immediate Actions

Organizations actively deploying or evaluating agentic AI systems should treat the Five Eyes "Careful Adoption of Agentic AI Services" guidance as an operational security baseline now, without waiting for domestic regulatory mandate. The guidance's principle of restricting agentic AI to low-risk, non-sensitive tasks until maturity is demonstrated is a defensible position that most organizations have not explicitly documented. Security teams should review current agentic deployments against each of the five risk categories—privilege, design and configuration, behavioral, structural, and accountability—and document where known gaps exist and what compensating controls are in place.

Privilege audits should specifically examine whether agentic service accounts hold persistent broad access or whether access is granted and revoked at the task level. Where persistent broad credentials exist for agentic systems, this should be treated as a priority remediation item regardless of whether any incident has been observed.

Short-Term Mitigations

Over the coming months, organizations should establish logging and audit trail requirements specifically designed for agentic decision chains, not merely for individual tool calls. A log that records what an agent called, but not why—the reasoning state and objective context at the moment of each decision—is insufficient for meaningful post-incident analysis. The OWASP Top 10 for Agentic Applications, published in December 2025, provides a technical vocabulary for the accountability deficit category that maps directly to audit trail design requirements [10].

Human-in-the-loop controls should be explicitly designed rather than assumed to exist through existing review processes. For agentic systems that can take irreversible actions—deleting records, sending external communications, executing financial transactions, modifying infrastructure configurations—automatic approval gates that pause execution and require human review should be treated as a baseline security requirement, not a performance tradeoff to be resolved in favor of throughput.

Strategic Considerations

At the strategic level, the governance divergence documented here creates a durable compliance planning challenge for organizations operating across jurisdictions. The appropriate response is to architect agentic AI governance programs against the conjunction of applicable standards rather than

the minimum: the EU AI Act for legal compliance, the Five Eyes guidance for security baseline, and emerging domestic standards—which may evolve as the political environment shifts—as supplements. Building governance programs against the most stringent applicable standard provides resilience against regulatory change in either direction.

Organizations should also engage with the voluntary standards process at NIST, even as its role has been reduced in scope, and with the OWASP Agentic Security Initiative and AIVSS working groups, which are producing the technical scaffolding that formal governance frameworks will eventually require [10]. The maturity of agentic AI security standards is early; organizations that contribute to their development will have an advantage when those standards become binding requirements.

CSA Resource Alignment

CSA's MAESTRO framework—Multi-Agent Environment, Security, Threat, Risk, and Outcome—published in February 2025, provides a seven-layer threat modeling architecture specifically designed for agentic AI systems [9]. MAESTRO's structure (Foundation Models, Data Operations, Agent Frameworks, Deployment and Infrastructure, Evaluation and Observability, Security and Compliance, Agent Ecosystem) maps directly onto the risk categories the Five Eyes guidance identifies. The privilege and accountability categories in the joint guidance correspond to MAESTRO's Deployment and Infrastructure and Evaluation and Observability layers; behavioral and structural risks correspond to the Agent Frameworks and Agent Ecosystem layers. Security teams implementing the Five Eyes recommendations should use MAESTRO as the threat model driving control selection.

CSA's AI Controls Matrix (AICM) addresses the governance gap at the organizational level by defining shared security responsibilities across model providers, application providers, orchestrated service providers, and AI customers—the same multi-tier principal structure through which agentic systems create accountability deficits. The AICM's Orchestrated Service Provider (OSP) guidelines are particularly relevant for multi-agent architectures where a primary deployer exercises only indirect control over agent behavior. CSA's STAR program provides the third-party assurance mechanism through which organizations can demonstrate conformance with AICM controls to customers and regulators, a capability whose value increases as regulatory expectations for agentic AI grow.

Zero Trust architecture principles—continuous verification, least-privilege access, assume-breach posture—align closely with the Five Eyes guidance's core technical recommendations and provide an existing organizational capability to anchor agentic AI security programs. Organizations with mature Zero

Trust implementations are better positioned to address the privilege and design categories of agentic AI risk, as these same underlying controls directly target the access-scope and configuration issues the guidance identifies.

References

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

[2] CISA. "CISA, U.S. and International Partners Release Guide on Secure Adoption of Agentic AI." CISA News, May 2026.

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

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

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

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

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

[8] TechPolicy.Press. "From Safety to Security: Renaming the U.S. AI Safety Institute Is Not Just Semantics." TechPolicy.Press, 2025.

[9] Huang, Ken. "Agentic AI Threat Modeling Framework: MAESTRO." Cloud Security Alliance, February 6, 2025.

[10] OWASP GenAI Security Project. "OWASP Top 10 for Agentic Applications." OWASP, December 9, 2025.

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

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

[13] Chatham House. "Breaking the Deadlock on AI Governance: Barriers to Global AI Governance." Chatham House, March 2026.

[14] NIST. "AI Risk Management Framework." National Institute of Standards and Technology, 2023 (ongoing revisions).

[15] CyberScoop. "CISA, NSA, Five Eyes Issue Guidance on Secure Deployment of AI Agents." CyberScoop, May 2026.

[16] UK House of Commons Library. "Artificial Intelligence: Regulation." House of Commons Library, 2025.

[17] Montreal AI Ethics Institute. "The Death of Canada's Artificial Intelligence and Data Act: What Happened and What's Next." Montreal AI Ethics Institute, 2025.

[18] BankInfoSecurity. "Australia Abandons Proposed Mandatory AI Rules in New Plan." BankInfoSecurity, December 2025.

[19] The Federal Register. "Executive Order 14179: Removing Barriers to American Leadership in Artificial Intelligence." Federal Register, January 31, 2025.

[20] Bloomberg. "Head of US AI Safety Institute to Leave as Trump Shifts Course." Bloomberg, February 4, 2025.

[21] European Commission AI Act Service Desk. "Timeline for the Implementation of the EU AI Act." European Commission, 2025.