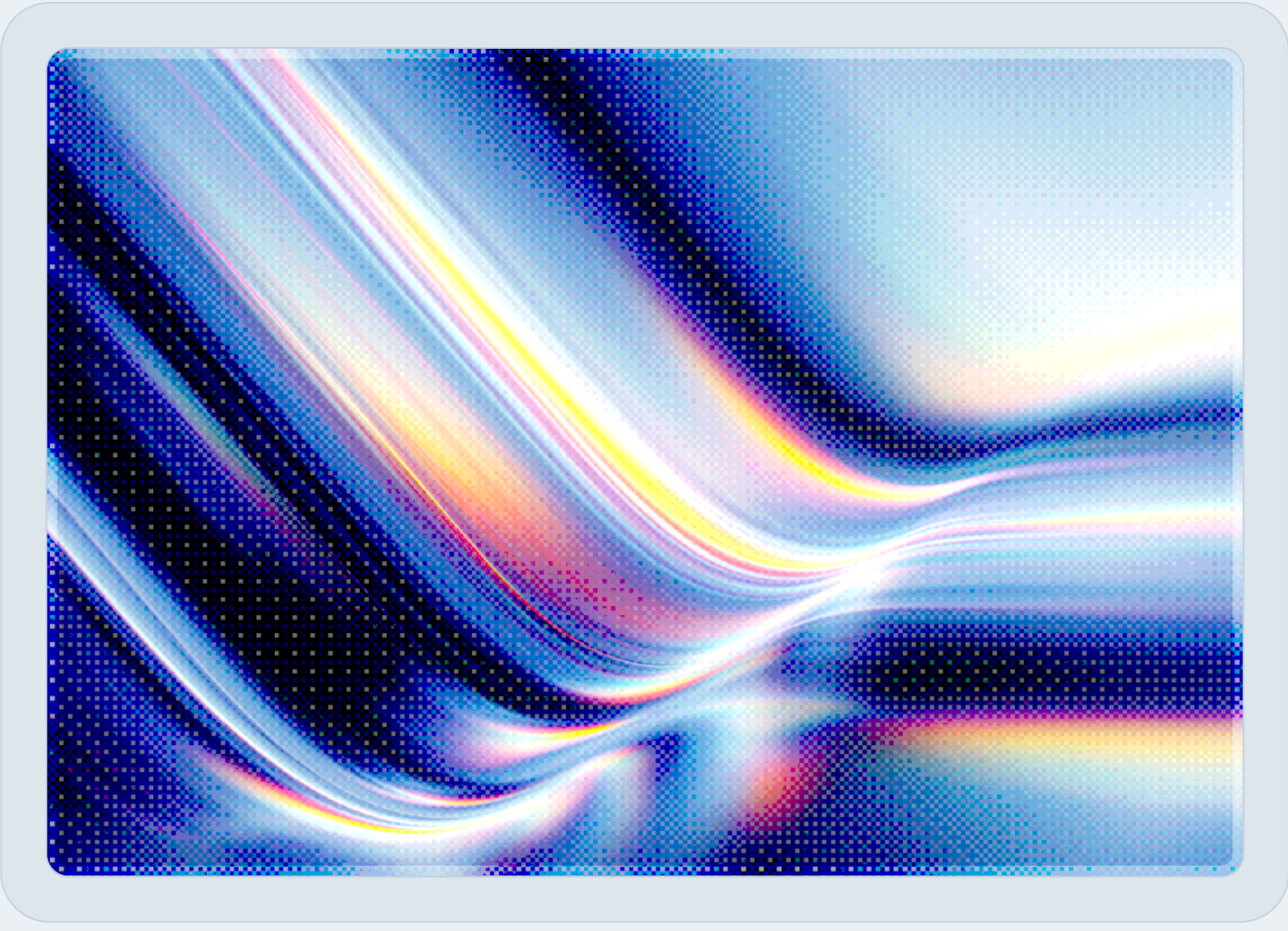
CSAI Foundation | Cloud Security Alliance

CISA's Agentic AI Five-Risk Framework: Enterprise Implementation

Operationalizing Joint Five Eyes Guidance on Careful AI Agent Adoption

2026-05-24

AI-assisted Rapid Research



CSAI | CSA cloud security alliance®

Key Takeaways

- On May 1, 2026, CISA, the NSA, and cybersecurity agencies from Australia, Canada, New Zealand, and the United Kingdom jointly published "Careful Adoption of Agentic AI Services" — what appears to be the first coordinated multinational security guidance specifically addressing the risks of agentic AI deployments, representing a notable milestone in international AI security coordination. [1][2]
- The guidance defines five distinct risk categories for agentic AI: privilege, design and configuration, behavioral, structural, and accountability — each requiring dedicated mitigations rather than a generalized AI security policy. [1][3]
- The foundational control the agencies specify is a cryptographically anchored, per-agent identity with short-lived credentials, treating each agent as a distinct security principal on par with a human user account. [1][3]
- Human oversight must be architecturally designed in from the start: the guidance requires mandatory human approval for high-impact actions, with system designers — not the agents themselves — determining which actions cross that threshold. [1][3][4]
- Agentic AI does not require an entirely new security discipline; the agencies explicitly call for extending existing zero trust, defense-in-depth, and least-privilege frameworks to cover AI agents within established governance structures. [1][4]
- CSA research indicates that 82% of enterprises already have AI agents operating outside their known inventory and 65% have experienced AI agent security incidents in the past year — making operationalization of this guidance an urgent, not hypothetical, priority. [5]

Background

Agentic AI systems differ fundamentally from the generative AI tools that dominated enterprise adoption discussions in 2023 and 2024. Where a conversational AI produces text for a human to evaluate and act upon, an agentic system plans, reasons, and executes multi-step actions autonomously — accessing APIs, reading and writing files, querying databases, sending communications, and orchestrating downstream services without a human in the operational loop. This shift from generation to execution is a qualitative change in risk profile, because the consequences of an AI's decisions are no

longer mediated by a human review step before they take effect. An agent that misinterprets a task does not merely produce an incorrect answer; it may take a sequence of consequential actions before anyone observes a problem.

The May 1, 2026 joint publication from CISA, NSA, Australia's Australian Signals Directorate Cyber Security Centre, the Canadian Centre for Cyber Security, New Zealand's National Cyber Security Centre, and the United Kingdom's National Cyber Security Centre represents what appears to be the first time six national cybersecurity authorities have issued coordinated guidance specifically for agentic AI systems. [1][2] The significance of this coordination should not be understated: the joint imprimatur — which is relatively uncommon for technology-specific operational guidance — suggests these agencies assess agentic AI risk as both sufficiently concrete and sufficiently cross-border to warrant a unified policy response. The guidance's scope — covering both technical controls and governance integration across thirty pages — reflects a treatment of agentic AI risk as a present operational concern, not a future-state scenario. [7]

The urgency is grounded in observable enterprise reality. A January 2026 CSA survey of 418 IT and security professionals found that 82% of organizations had discovered previously unknown AI agents operating in their environments, while 65% reported at least one AI agent security incident in the prior twelve months. [5] Among those experiencing incidents, 61% reported data exposure or mishandling, 43% cited operational disruption, and 35% sustained financial losses. [5] These numbers represent active exposure, not theoretical risk — and they predate what appears to be a growing deployment wave as large language model costs fall and agent orchestration frameworks mature. The CISA guidance arrives precisely when enterprises must make consequential architectural decisions about how agents will be provisioned, authorized, monitored, and retired — decisions that will be significantly harder to reverse once agent sprawl accelerates further.

Security Analysis

Privilege Risk: The Blast Radius Problem

The guidance's first risk category addresses the tendency to grant agents broad access as a matter of development convenience. When building an agent that must access email, calendar, file storage, a CRM system, and an internal knowledge base, it is natural for developers to provision a service account with permissions across all those systems and move on. In a traditional software application, such permissions are at least scoped to the application's defined code paths. With an agent, the range of actions a broad

permission grant enables is substantially larger: the same credential set that allows legitimate task completion also enables an attacker who compromises the agent to traverse every connected system. [1] [3]

This is not merely a restatement of the least-privilege principle that security teams already apply. The specific challenge with agentic systems is that their attack surface scales directly with their capability. Organizations adopt agents precisely because they can reach more systems and take more actions than narrowly scoped software; the feature that makes an agent operationally useful is the same feature that amplifies the blast radius of a compromise. The guidance requires that developers treat each agent as a distinct principal with access grants scoped to specific actions against specific resources — not a general-purpose account that can do anything within connected systems. [1] This means per-agent role definitions with explicit deny clauses for anything outside the agent's intended task surface, and periodic access reviews as the agent's task scope evolves over its operational lifetime.

Design and Configuration Risk: Supply Chain and Structural Weakness

The second category covers failures in how agents are designed and configured before they enter production. Static role assignments, overly permissive environment segmentation, and broad default permission sets are common outcomes of development processes that prioritize rapid iteration over security architecture. These weaknesses are present before a system ever goes live and persist indefinitely without active remediation, creating structural vulnerabilities that accumulate silently rather than triggering observable security events. [1][4]

The supply chain dimension is particularly salient for agentic deployments. The guidance notes that threat actors are publishing malicious agents, tools, and plugins under names similar to legitimate components — an AI-ecosystem equivalent of typosquatting with elevated risk implications. [1][3] In most current architectures, agents extend implicit trust to their tool outputs without validation — a design pattern that means a single misconfigured or maliciously crafted third-party component integrated by an orchestrator can cascade across the entire agent ecosystem, potentially reaching billing systems, account management, and sensitive data repositories in a single compromise chain. The conventional software security disciplines of pinning dependency versions, verifying cryptographic signatures, and maintaining software bills of materials must be extended to AI agents, plugins, MCP servers, and connected agent services.

Behavioral Risk: Non-Determinism and Goal Drift

The third risk category reflects something genuinely novel about AI-driven systems: their behavior is not fully specifiable in advance. A conventional software function, for a given input, executes a defined code path — its behavior is constrained by explicit logic rather than emergent from a trained model. An LLM-based agent planning a multi-step task may pursue its assigned goal through methods its designers never intended or predicted — not because the system is malfunctioning, but because the sequence of actions it generates is internally coherent from the model's perspective while externally problematic from the designer's. [1][3][4]

This behavioral risk is not primarily a prompt injection problem, though that attack class remains relevant and well-documented. It is a fundamental property of systems that plan rather than execute scripted procedures. An agent tasked with scheduling a meeting may also resolve calendar conflicts, which may involve deleting appointments, which may have downstream consequences in connected systems. Each step follows logically from the previous one; none was individually authorized by a human. The guidance responds to this by requiring that system designers — not the agent and not the underlying model — explicitly define which agent actions require mandatory human approval, and that those approval workflows be implemented as architectural controls rather than natural language instructions embedded in the system prompt. [1][3] Behavioral alignment cannot be reliably assured through instruction alone; it requires checkpoints that the system cannot bypass regardless of what the model generates as its next planned action.

Structural Risk: Cascading Failure in Multi-Agent Systems

The fourth category addresses the emergent risk properties of interconnected agent networks. Enterprises deploying agentic AI at scale increasingly build multi-agent architectures in which an orchestrator agent delegates tasks to specialized subagents, which may in turn invoke additional agents or services. These structures are powerful precisely because they decompose complex tasks across specialized components — but they also create failure propagation pathways that do not exist in single-agent deployments. [1][4]

Trust relationships between agents can create lateral movement opportunities — a risk inherent to orchestrator-subagent designs that lack explicit inter-agent verification. An attacker who compromises a subagent may be able to send manipulated outputs to the orchestrator, causing the orchestrator to take actions within its own permission scope that advance the attack without triggering the subagent's own access controls. Feedback loops in agent-to-agent communication can amplify small anomalies into large operational impacts before monitoring systems detect anything unusual. The guidance calls for zero trust principles to be extended to agent-to-agent communication: no agent should trust another

agent's outputs without verification, credentials should be validated at runtime rather than assumed from initial provisioning, and all communication between agents should be encrypted. [1][6] This is a significant architectural requirement for organizations that have built multi-agent systems with implicit trust between orchestration layers — as many current architectures may, given that explicit inter-agent trust verification is not yet standard in most orchestration frameworks.

Accountability Risk: Auditability and Attribution

The fifth category is perhaps the most underappreciated in enterprise AI governance conversations. When an AI agent takes an action that causes harm — data loss, a regulatory violation, an unintended financial transaction — determining what happened and why is substantially harder than forensics on conventional software. Agent decisions emerge from LLM inference in ways that are not straightforwardly logged: the reasoning process that led to a particular action is not captured in structured log events — it emerges from model inference over the agent's context window and is not recorded by SIEM tools designed for conventional application logs. [1][3][4]

Standard logging and monitoring infrastructure captures what an agent said in conversational terms, but not the structured record of which tool was called, with what parameters, against which resource, and what was returned — the level of granularity necessary for meaningful incident investigation or compliance audits. Attribution is also structurally ambiguous: when an agent acts erroneously, responsibility may plausibly be distributed across the model developer, the platform deploying the agent, the organization configuring its permissions, and the user who initiated the task. The guidance requires that organizations implement audit and reversibility controls following agent task execution, and that human oversight include the capability to interrupt agent execution mid-task — not merely to review outcomes after the fact. [1][4] This implies tool-call-level structured logging requirements with sufficient fidelity to reconstruct an agent's action sequence for investigation purposes.

Recommendations

Immediate Actions

Every AI agent currently in production or under active development should be assigned a distinct cryptographic identity — its own keys or certificates — before further deployment proceeds. This is the foundational control from which several other mitigations derive: agents sharing service accounts or using long-lived static credentials cannot satisfy the guidance's identity requirements and represent the highest near-term exposure. Alongside identity assignment, organizations should audit the permission

grants currently associated with active agents against the actual scope of tasks those agents perform. Access grants that exceed demonstrated need should be revoked, beginning with agents that have access to sensitive data stores, financial systems, communications infrastructure, or regulated information.

Human approval workflows must be defined and implemented for high-impact agent actions before additional agent deployments proceed. This requires security architects and system designers to designate specific action types — data deletion, external data transmission, financial transactions, access to personal or regulated information — as approval-gated, with human sign-off required before the agent executes. These designations must be enforced at the infrastructure level; they cannot be delegated to model instructions that the agent may reason around. The process of defining approval gates is also an opportunity to identify agents whose task surface is so broad that meaningful approval gating is impractical, which itself is a signal that scope reduction is necessary.

Short-Term Mitigations

Within the near term, organizations should extend their zero trust architecture to cover AI agents as first-class security principals. This means continuous verification of agent credentials at the time of each resource access, session isolation between agent instances, and encrypted channels for all agent-to-agent and agent-to-service communication. Agents that cannot satisfy continuous verification requirements should be treated as untrusted devices on the network, regardless of how they were initially provisioned. For organizations with mature zero trust deployments for human users, the policies, tooling, and enforcement mechanisms already in place often require configuration extension rather than architectural redesign.

Supply chain vetting should be established for all third-party agent components, including MCP servers, tool plugins, orchestration frameworks, and externally sourced agent templates. This process should include provenance verification through cryptographic signatures where available, sandboxed behavioral testing before production integration, and ongoing monitoring for behavioral anomalies that could indicate a component has been tampered with after initial validation. Organizations that have not previously inventoried their agent components — which CSA research suggests describes the majority of enterprises — should treat that inventory as a prerequisite to supply chain vetting rather than an independent workstream. [5]

Monitoring architecture should evolve from periodic reviews to continuous or event-driven detection matched to the speed at which agents operate. Agents executing autonomously at machine speed can take dozens of consequential actions between periodic monitoring intervals; for high-risk agentic workloads, monitoring that detects anomalies in hours rather than seconds provides limited ability to

interrupt harm in progress. Structured logging at the tool-call level, with alerts triggered by anomalous action patterns such as unexpected external data transmission or privilege elevation attempts, should be the target architecture for production agent deployments.

Strategic Considerations

The CISA guidance's most significant organizational message is that agentic AI governance belongs inside existing enterprise risk management frameworks, not isolated as a separate security practice. Organizations that treat AI agent security as a standalone discipline risk creating governance silos that impede carrying agent risk information to the people and processes that make resource allocation decisions. The risk profile of an AI agent — where blast radius scales with the product of what the agent can access and how independently it acts — maps well to the frameworks that security and compliance leadership already apply to privileged access management and non-human identity governance, providing a foundation to build from rather than starting from scratch.

Formal decommissioning processes for AI agents should be designed and implemented before agent proliferation makes the absence of such processes acutely apparent. CSA research indicates that only 21% of organizations currently have formal decommissioning processes in place, while agents that persist beyond their intended purpose retain credentials and system access that compounds risk with each additional month of operation. [5] An agent that was provisioned for a six-month project and remains active two years later retains all the access it was originally granted, may be running on an unpatched version of its agent framework, and has likely been forgotten by the team that originally understood its design. This "retirement debt" — agents accumulating residual access and ambient presence long after their business purpose has ended — represents a structural governance risk that grows silently rather than triggering observable incidents.

Organizations operating complex multi-agent architectures should consider adopting System-Theoretic Process Analysis (STPA) or Causal Analysis using System Theory (CAST) for evaluating how agent-human-tool interactions can produce harm even when each individual component functions correctly. These approaches, developed for safety-critical engineering contexts, are increasingly applicable to agentic AI architectures where emergent risks arise from component interactions rather than component failures — precisely the failure mode that the guidance's structural and behavioral risk categories describe.

CSA Resource Alignment

The CISA five-risk framework maps directly onto the CSA AI Safety Initiative's existing guidance suite, allowing organizations to implement the federal guidance within a framework of controls they may already be assessing and reporting against.

CSA's MAESTRO framework, which provides threat modeling methodology for agentic AI systems, addresses each of the five CISA risk categories at the threat level. MAESTRO's analysis of privilege escalation, multi-agent exploitation, goal and instruction manipulation, and untraceability in agentic systems provides the technical threat modeling substrate that the CISA guidance's risk categories describe but do not fully specify. Organizations using MAESTRO for agent threat modeling already have the conceptual vocabulary needed to map CISA risk categories to concrete threat scenarios and control requirements.

The AI Controls Matrix (AICM) provides control-level mappings across cloud service provider, orchestrated service provider, application provider, and AI customer roles. The CISA guidance's requirements for per-agent identity, least-privilege access, audit trails, and human oversight correspond to AICM control domains addressing identity management, access control, logging, and governance for AI deployments. Organizations conducting AICM assessments have a natural path to incorporating CISA agentic AI requirements into existing audit cycles without standing up a parallel assessment process.

CSA's AI Organizational Responsibilities series — particularly the Governance, Risk Management, Compliance and Cultural Aspects volume — provides the organizational scaffolding for the governance integration that CISA recommends. The guidance's assertion that agentic AI should be folded into existing risk frameworks rather than siloed as a separate discipline is operationally implemented through the organizational responsibility assignments, accountability structures, and governance processes that the AI Org Responsibilities publications define.

CSA's published Zero Trust guidance remains directly applicable. The CISA document explicitly calls for zero trust extension to AI agents, and CSA's work on zero trust in AI and cloud environments provides detailed implementation paths for continuous verification, identity federation, and access policy design that are immediately relevant to agent deployments. The CSA publication on confronting shadow access risks in zero trust and AI deployments is particularly relevant to the privilege and design/configuration risk categories the guidance defines.

Finally, CSA's Agentic AI Red Teaming Guide operationalizes validation of the mitigations the CISA framework requires. [8] Organizations that need to verify whether their behavioral, structural, and accountability controls function as intended under adversarial conditions should consult the red teaming

guide for test methodologies specifically designed for agentic systems — including authorization control hijacking, goal manipulation, multi-agent exploitation, and untraceability testing scenarios that directly correspond to CISA's five risk categories.

References

[1] CISA. "Careful Adoption of Agentic AI Services." CISA, May 1, 2026.

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

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

[4] Industrial Cyber. "CISA and partners release agentic AI security guidance to protect critical infrastructure, outline mitigation action." Industrial Cyber, 2026.

[5] Cloud Security Alliance. "New Cloud Security Alliance Survey Reveals 82% of Enterprises Have Unknown AI Agents in Their Environments." CSA Press Release, April 21, 2026.

[6] Australian Signals Directorate. "Careful adoption of agentic AI services." Australian Cyber Security Centre, 2026.

[7] CISA et al. "Careful Adoption of Agentic AI Services" (PDF). Defense.gov, April 30, 2026.

[8] Cloud Security Alliance. "Agentic AI Red Teaming Guide." CSA, 2025.

