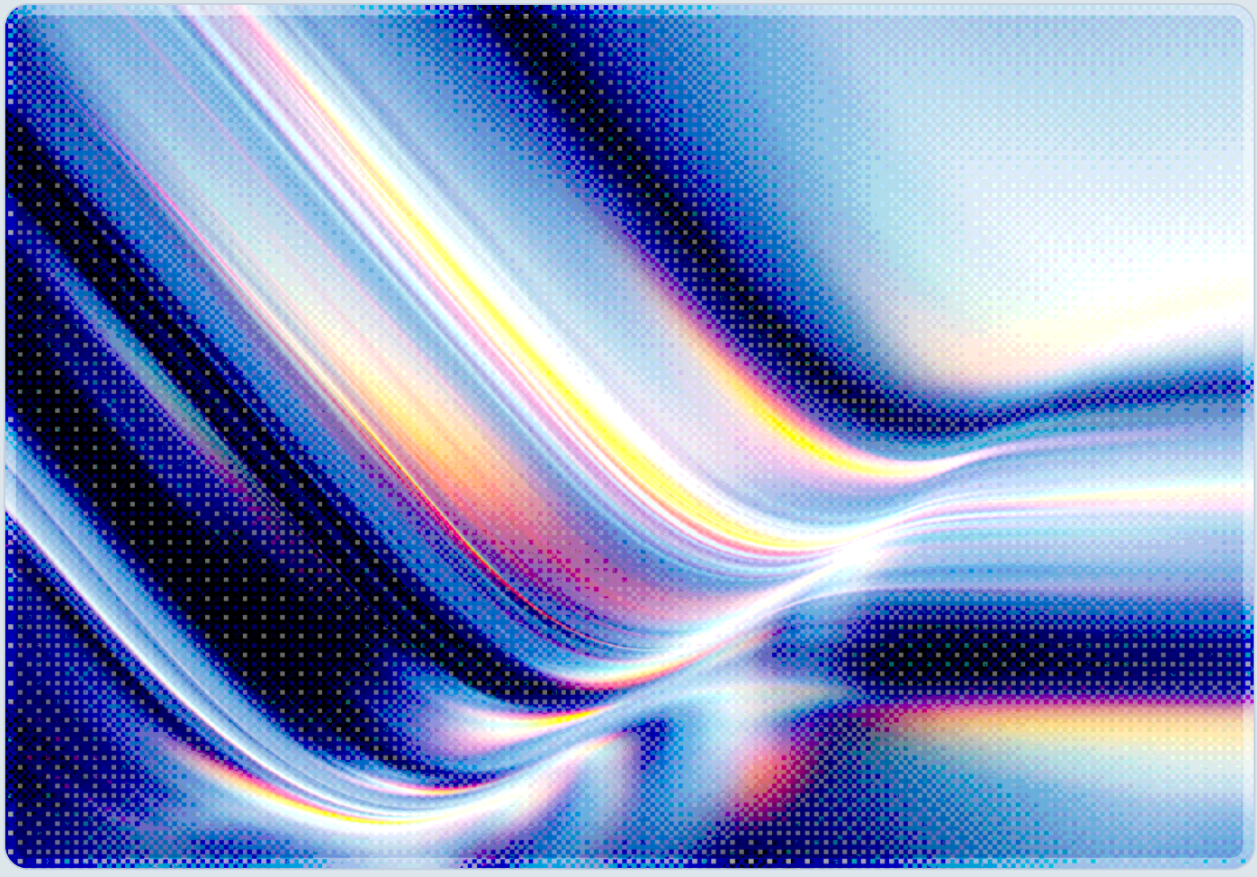
CSAI Foundation | Cloud Security Alliance

OWASP's Agentic AI Maturity Model: A CISO Guide

Reading Version 2.01 of the State of Agentic AI Security and Governance Report

2026-07-03

AI-assisted Rapid Research



CSAI

CSA cloud security alliance®

Key Takeaways

On June 1, 2026, the OWASP GenAI Security Project published version 2.01 of its "State of Agentic AI Security and Governance" report, a substantial update to the July 2025 first edition that reframes agentic AI risk around evidence rather than speculation [1]. Where v1.0 treated agentic threats as a portfolio of plausible future problems, v2.01 draws on a year in which the OWASP Top 10 for Agentic Applications was published, a body of confirmed incidents accumulated behind nearly every one of its ten risk categories, and a regulatory environment for agent-driven harm began to take concrete shape [1][2]. The report organizes itself around three findings consistent with the broader shift in agentic-AI risk reporting over the past year: the threats it describes are no longer hypothetical, safety and security failures converge at the point where an organization deploys an agent, and governance capability is falling further behind deployment speed rather than catching up [1].

The centerpiece of the update, and the reason it is directly useful to CISOs rather than only to practitioners, is a two-dimensional Enterprise Adoption Maturity Model. One axis, the Adoption Tier, classifies what an organization is actually running, from unmanaged "Shadow AI" through vendor-embedded assistants to federated, cross-organizational agent networks. The second axis, Governance Maturity, scores an organization's capability to oversee whatever it is running, from ad hoc awareness to adaptive, telemetry-driven control [1]. Crossing the two produces a matrix that tells a security leader, for a specific class of agent deployment, whether current governance is adequate, insufficient, or a decision the organization should not be making at all. The scale of adoption underscores the stakes: OWASP cites Andreessen Horowitz's April 2026 analysis finding that 29% of the Fortune 500 and roughly 19% of the Global 2000 are already live, contracted customers of a leading AI vendor, a penetration rate reached roughly three years after ChatGPT's public launch, and one that by construction excludes the unmanaged usage the maturity model is built to surface [1][3].

Background

The OWASP GenAI Security Project is an OWASP Flagship initiative that reports more than 600 contributors from 18 countries, and the Agentic Security Initiative (ASI) is the stream within it responsible for agent-specific guidance [1]. Since the first edition of this report in July 2025, arguably the initiative's most consequential release has been the OWASP Top 10 for Agentic Applications, published in December 2025 after input the project describes as coming from more than 100 security researchers

and practitioners, which established the first standardized taxonomy of agentic risk: from Agent Goal Hijack (ASI01) and Tool Misuse (ASI02) through Identity and Privilege Abuse (ASI03), Agentic Supply Chain Vulnerabilities (ASI04), Unexpected Code Execution (ASI05), Memory and Context Poisoning (ASI06), Insecure Inter-Agent Communication (ASI07), Cascading Failures (ASI08), Human-Agent Trust Exploitation (ASI09), and Rogue Agents (ASI10) [2]. Version 2.01 of the governance report uses that taxonomy as its organizing spine, and the update's authors note that almost every one of the ten categories now has at least one confirmed real-world incident behind it, a change from the largely architectural concerns the first edition catalogued [1].

The report frames its central argument as a collapse of a distinction security teams have historically relied on: that AI safety, meaning harm arising from a system's normal operation, and AI security, meaning harm arising from a trust boundary being crossed, could be owned by separate teams with separate methods. Once a system can act autonomously against production infrastructure, OWASP argues, that separation no longer holds at what it calls the deployment layer, the architectural decisions, permissions, and operational controls owned by the organization running the agent rather than the model provider that built it [1]. A coding agent with an overly permissive sandbox that receives a prompt injection through a cloned repository, and a coding agent that deletes a production database because its own reliability failed under an ambiguous instruction, produce the same outcome through different mechanisms, and the report's position is that an organization needs one incident response and governance structure capable of catching both, not two [1]. Regulators have converged on a related premise: the report notes that Europe's Digital Operational Resilience Act requires major incident notification within four hours, the NIS2 Directive requires a 24-hour early warning, New York's RAISE Act sets a 72-hour frontier-incident reporting window, and California's SB 53 requires disclosure within 15 days, all of which assume continuous monitoring capability rather than periodic audit [1].

To support that argument, v2.01 introduces a revised Agents Taxonomy built on three independent dimensions rather than a single classification. Agent Type describes what an agent does and where it operates, spanning enterprise agents serving internal users, coding agents that generate and modify code, client-facing agents interacting directly with customers, personal agents running with a user's own device permissions, and infrastructure and operations agents managing cloud resources and pipelines. Implementation Pattern describes how an agent is built, whether through an orchestration framework such as LangGraph or CrewAI, a lightweight library built from SDKs, or a platform-native, low-code environment such as Copilot Studio, a distinction that matters because it determines how visible the agent is to organizational audit. Composition Pattern describes how agents are arranged, from a single agent calling tools, to tightly coupled multi-agent systems, to loosely coupled distributed agent chains communicating over protocols like Google's A2A or Anthropic's MCP, to agent-spawning architectures in which a parent agent dynamically creates ephemeral sub-agents. Autonomy level cuts across all three as the report's fourth, risk-scaling dimension, ranging from supervised operation where a human approves

each action, through semi-autonomous operation with human review of flagged items, to fully autonomous operation where kill switches, budget limits, and a dedicated agent identity become the primary safeguards rather than a human in the loop [1].

Security Analysis

The Adoption Tier scale is where the maturity model becomes concrete. It defines nine tiers, AT0 through AT8, ordered by escalating trust boundary and autonomy characteristics rather than by organizational intent, since the lowest tier is explicitly not a deliberate deployment choice.

Tier	Name	Core Characteristic	Representative Examples
AT0	Shadow AI	No organizational awareness or approval; users self-adopt tools outside governance	Personal ChatGPT/Gemini/Claude use on corporate data, unapproved browser AI extensions
AT1	Vendor Embedded Assistant	Fully vendor-controlled; consumed, not built	Microsoft 365 Copilot, GitHub Copilot, Salesforce Einstein
AT2	Platform Integrated	AI-native platform using organizational data; cannot execute arbitrary code	Custom GPTs, Amazon Q Business, Google Vertex AI Agents
AT3	Citizen Developer Agent	Low-code/no-code platform; user configures flows against real organizational data	Power Automate + AI Builder, Copilot Studio, Zapier Central AI
AT4	Code Executing Agent	Generates and executes code with local or cloud privileges	Open Interpreter, coding assistants, GitHub Copilot Workspace
AT5	Custom In-House Agent	Organization built it and controls identity, tools, and boundaries	LangChain/LangGraph custom agents, in-house RAG pipelines

Tier	Name	Core Characteristic	Representative Examples
AT6	Externally Extended Agent	Connects to external tools and services across trust boundaries	Agents with MCP servers, agents calling third-party APIs
AT7	Multi-Agent Orchestration	Multiple agents coordinate within the organization	CrewAI workflows, AutoGen multi-agent teams
AT8	Federated / Cross-Boundary	Agents operate across organizational boundaries	Cross-org supply chain agents, multi-tenant agent marketplaces

The second axis, Governance Maturity, scores organizational capability on a 0-to-4 scale independent of what is deployed. Level 0, Unaware and Ad Hoc, describes an organization with no formal recognition of agentic risk beyond traditional AI concerns and only informal, minimal oversight. Level 1, Experimentation without Guardrails, describes pilot projects that lack defined autonomy limits or escalation criteria, governed by generic policy rather than continuous monitoring. Level 2, Policy-Defined and Human-in-the-Loop, introduces published governance policy, mandatory human review for high-impact decisions, a named accountable executive, and established logging and AI bill-of-materials practices, though monitoring remains periodic. Level 3, Integrated and Continuous Oversight, treats agentic AI as critical infrastructure with real-time anomaly detection, working kill switches, and governance enforced as machine-readable policy. Level 4, Adaptive and Self-Regulating Governance, describes organizations whose telemetry, red-team findings, and regulatory inputs automatically tune guardrails, with cryptographic agent identity and live, tamper-evident audit trails supporting deployment across jurisdictions [1].

One of the report's most actionable additions is what it calls the Maturity Level by Adoption Tier posture matrix, which crosses the two scales and labels the resulting governance posture for each combination, from well-governed through adequate, high-exposure, critical gap, and, at the extreme, do-not-deploy. An organization at Governance Level 0 or 1 running AT6 or AT7 externally extended or multi-agent systems sits in a cell the report marks as a critical gap: the full ASI risk surface is exposed without supply-chain verification or agent-to-agent authentication in place. Federated, cross-boundary deployment at AT8 is marked do-not-deploy below Governance Level 3 under any circumstances, since federated trust requires the continuous oversight and cryptographic identity that only the two highest maturity levels provide. AT0 Shadow AI is treated as a special case throughout: because it is unmanaged by definition, it cannot be governed into a better posture, only discovered and eliminated by moving it into a managed tier or blocking it outright, and the report instructs organizations to assume it exists until proven otherwise through network telemetry, data-loss-prevention signals, and employee surveys [1].

The model's operational value comes with a real adoption cost that the report does not dwell on. Cross-referencing nine Adoption Tiers against five Governance Maturity levels produces forty-five distinct posture cells, and populating that matrix accurately requires an agent inventory most organizations do not yet have the tooling to produce, particularly for tiers such as AT3 citizen-developer automations and AT4 code-executing agents that individual business units can stand up without central visibility. The framework is also new: this is the first governance-maturity edition of the report, and its tier and level definitions have not yet been tested against a full audit cycle. None of that diminishes the model's usefulness as a starting structure, but security leaders should treat the initial classification exercise as a multi-quarter discovery effort rather than a one-time assessment.

The threat data that follows suggests the instruction to assume Shadow AI exists is more than precautionary. The report's revised threat analysis documents a shift from architectural concern to active exploitation over the twelve months since the first edition, concentrated in exactly the tiers the maturity model flags as highest risk. Prompt injection now maps to six of the ten ASI categories and functions as the primary delivery mechanism across the board, because large language models process system instructions, user input, and retrieved content as a single undifferentiated token stream with no consistently enforced privilege boundary between them [1]. The agentic supply chain, corresponding to the AT6 externally extended tier, moved from theoretical concern to the highest-volume incident category in the same period: researchers documented postmark-mcp, the first confirmed malicious MCP server, which shipped fifteen clean versions before adding a single line of data-exfiltration code, and a critical remote-code-execution flaw, CVE-2025-6514, was disclosed in mcp-remote, infrastructure used by hundreds of thousands of developers, carrying a CVSS score of 9.6 [1][4]. Coding agents at the AT4 and AT5 tiers showed a parallel pattern: CVE-2026-22708, disclosed against the Cursor AI code editor, demonstrated that an attacker able to influence an agent's instructions could cause shell built-in commands to silently bypass an allowlist that had already approved the surrounding command, turning an approval control into part of the attack path rather than a defense against it [1][5]. The report also describes a March 2026 campaign it attributes to an autonomous, deliberately weaponized bot that exploited GitHub Actions misconfigurations, compromised a security vendor's own repository through personal-access-token theft, and published backdoored packages directly to PyPI without further human direction. Independent public reporting on the episode is less settled: it identifies the initial GitHub Actions exploitation and the later PyPI-publishing phase as parts of a structured, multi-phase campaign whose later stages carry signs of human orchestration rather than a single uninterrupted autonomous chain end to end. Either characterization is consistent with the broader pattern the report draws from the incident: the AT6 through AT8 tiers now face adversarial automation on both sides of the boundary.

Recommendations

Immediate Actions

Security leaders should treat AT0 discovery as the non-negotiable starting point the report recommends, since no other step in the maturity model is meaningful if unmanaged agentic usage remains uninventoried. That means using network telemetry, data-loss-prevention tooling, and structured employee surveys to surface personal AI accounts, browser extensions, and locally run models operating against corporate data, on the assumption that this usage exists in essentially every organization until proven otherwise. In parallel, security teams should conduct an honest self-assessment of current Governance Maturity Level using the 0-to-4 criteria, and build or complete a registry of every deployed agent, tagging each by its Adoption Tier, since most organizations will find they operate across four or five tiers simultaneously, from a vendor-embedded assistant to a citizen-developer automation to an unrecognized custom coding agent.

Short-Term Mitigations

Once agents are classified by tier, organizations should check each against the Maturity Level by Adoption Tier matrix and treat any cell the report marks as a critical gap or do-not-deploy as a forcing function: either raise governance maturity to the required level or reduce the deployment's autonomy and trust boundary to match current capability. Control investment should be prioritized by the dominant ASI risk classes for an organization's highest active tier, with AT1 through AT4 deployments focused on Agent Goal Hijack, Unexpected Code Execution, and Memory and Context Poisoning, and AT6 and higher deployments requiring coverage of the full ASI01 through ASI10 surface with particular attention to supply chain verification, inter-agent authentication, and cascading-failure containment. Concretely, this means enforcing branch protection and required-reviewer policies at the repository layer rather than in agent configuration that the agent itself can influence, rotating any default credentials or signing keys before connecting an agent to external tools, and instituting human-in-the-loop approval for any agent session that combines access to private data, exposure to untrusted content, and the ability to communicate externally, the three-property combination the report's threat analysis identifies as sufficient for a single prompt injection to complete an entire attack chain.

Strategic Considerations

The report's structural argument, that safety and security cannot remain organizationally separate once a system acts autonomously against production systems, has a direct implication for how security functions should be resourced: incident response, threat modeling, and governance ownership for agentic deployments should sit under a single accountable structure rather than split between an AI safety function and a security operations function that investigate the same telemetry for different causes. Governance investment should also be sequenced deliberately, moving from static, document-driven oversight toward adaptive, telemetry-backed control loops as deployment complexity increases, rather than attempting to reach Level 4 adaptive governance uniformly before any tier-appropriate deployment proceeds. Finally, organizations approaching AT7 multi-agent orchestration or AT8 federated deployments should treat Governance Level 3 continuous oversight, including working kill switches and cryptographic agent identity, as a precondition rather than a target to reach after deployment, given the report's explicit position that federated trust arrangements are not governable at lower maturity levels.

CSA Resource Alignment

OWASP's report explicitly credits CSA's MAESTRO threat-modeling framework as one of several independent efforts, alongside AWS's architecture scoping matrix, NVIDIA and Lakera's safety framework, and Google's A2A protocol design, that converged on the same conclusion: agentic risk cannot be reduced to a single layer but emerges from the interaction between a model's reasoning capability, the tools it can invoke, accumulated memory and context, and trust relationships between cooperating agents [1]. Organizations applying OWASP's Adoption Tier and Governance Maturity model alongside MAESTRO can use MAESTRO's seven-layer decomposition, Foundation Models, Data Operations, Agent Frameworks, Deployment and Infrastructure, Evaluation and Observability, Security and Compliance, and Agent Ecosystem, to identify which specific layer is under-governed at a given adoption tier, giving security teams a way to translate a matrix cell marked as a gap into a specific architectural remediation [6]. The credential rotation, identity scoping, and supply-chain verification controls the maturity model calls for at the AT4 through AT8 tiers map directly onto control domains in CSA's AI Controls Matrix, which organizations can use to structure a defensible baseline for non-human identity management and third-party AI service integration as they move agents up the governance maturity scale [7]. The report's insistence that federated, cross-boundary agent deployments require mutual attestation and cryptographic trust before they are governable at all is a direct application of CSA's Zero Trust guidance, which argues against implicit trust between network-adjacent services, a principle the OWASP model applies specifically to agent-to-agent trust boundaries at the AT6 through

AT8 tiers [8]. Finally, the underlying ASI01 through ASI10 taxonomy that structures both the threat analysis and the maturity model's risk prioritization is the OWASP Top 10 for Agentic Applications itself, and CSA's own agentic AI research has consistently used that taxonomy as a shared reference point for describing incidents, reinforcing the case for a common vocabulary between the two organizations' guidance [2].

References

[1] OWASP GenAI Security Project. "State of Agentic AI Security and Governance, Version 2.01." OWASP, June 2026.

[2] OWASP GenAI Security Project. "OWASP Top 10 for Agentic Applications for 2026." OWASP, December 9, 2025.

[3] Andreessen Horowitz. "Where Enterprises Are Actually Adopting AI." a16z, April 8, 2026.

[4] National Institute of Standards and Technology. "CVE-2025-6514 Detail." National Vulnerability Database.

[5] National Institute of Standards and Technology. "CVE-2026-22708 Detail." National Vulnerability Database.

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

[7] Cloud Security Alliance. "AI Controls Matrix (AICM) v1.1." CSA.

[8] Cloud Security Alliance. "Zero Trust Guiding Principles." CSA.