OWASP | GenAI SECURITY PROJECT

State of Agentic AI Security and Governance

Version 2.01
June 2026

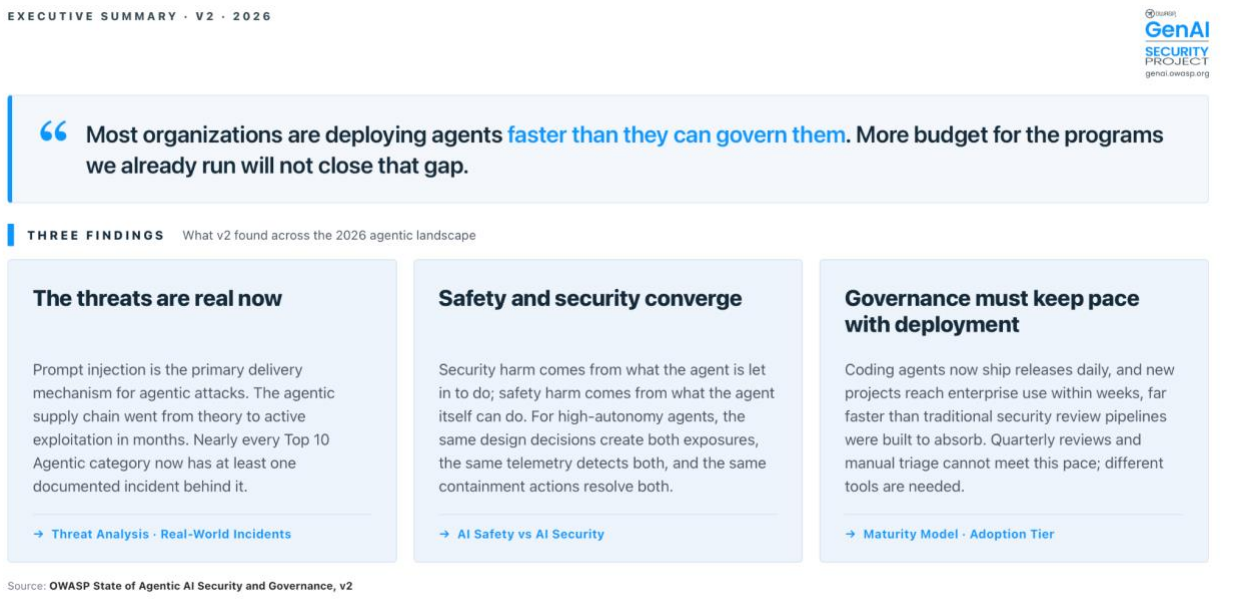
Our mission is to provide actionable insights into the security challenges of Agentic AI, helping organizations develop, deploy, and govern these systems responsibly. We empower security professionals with the tools and knowledge needed to understand the evolving ecosystem of tools and emerging regulations on AI, mitigate risks, ensure compliance, and drive safe AI innovation.

License and Usage

Table of Content

Executive Summary

v1.0 of this report (July 2025) framed agentic risks as a portfolio of plausible threats, surveyed an early ecosystem, and called for governance to keep pace. In the year since, adoption has accelerated, the OWASP Top 10 for Agentic Applications has been published, and a regulatory environment around agent-driven harm has begun to take shape. v2 reads that year of evidence and is organized around three findings.



The threats are real now. What was a list of architectural concerns in 2025 now has production incidents, vendor advisories, and CVEs attached to almost every entry. The Real-World Incidents and Exploits Tracker in this edition curates that body of evidence, and the Threat Analysis chapter places each pattern in the context of where agentic capability was expanding when the break happened. The threat model is no longer hypothetical, and the design conversations that follow from it are no longer optional.

Safety and security converge at the deployment layer. For most of software's history, safety has been an engineering concern and security has been an adversarial one, owned by different teams with different methods. Agentic systems blur this picture when they act with broad autonomy and tool access. We argue that at the *deployment layer* — the architectural decisions, configurations, permissions, and operational controls owned by the deploying organization — the two categories cannot be operationally separated.We argue that at the deployment layer the two categories cannot be operationally separated. Model-level safety remains the provider's responsibility, but once an agent is acting on production systems the same controls

govern both kinds of harm and the same investigation surfaces both kinds of cause. The organisational implication is direct: AI Safety and AI Security cannot continue as parallel functions.

Governance must keep pace with deployment. Regulators have accepted the premise that agents can cause harm faster than human review can keep up. DORA's four-hour notification, NIS2's 24-hour early warning, NY RAISE's 72-hour frontier reporting, and CA SB 53's 15-day window all assume continuous oversight rather than periodic audit. The work that remains is different in kind, requiring live monitoring of agent behavior, baselines that flag drift, automated incident routing, and stop mechanisms that operate in seconds rather than days.

What is new in State of Agentic AI Security and Governance v2

This edition introduces a revised Threat Analysis grounded in documented incidents, a new chapter on AI Safety vs AI Security, and a Real-World Incidents and Exploits Tracker tied to the OWASP Top 10 for Agentic Applications. A new Enterprise Adoption Maturity Model assesses governance capability against deployment complexity, with each chapter aligned to the relevant Top 10 for Agentic categories. Agent Identity and Non-Human Identity is elevated to a chapter treating identity as the new control plane, joined by a new chapter on AI SBOM and Supply Chain Provenance. The Agents Taxonomy is revised across three independent dimensions (type, implementation, and composition with autonomy as the cross-cutting dimension), the ecosystem view draws on telemetry from 53 tracked agentic projects, and the regulatory landscape now covers 42 instruments across 10 jurisdictions.

Where to start

The practical starting point is to identify the most advanced agents you are running today, then either raise governance maturity to match or reduce the deployment tier. Particular attention is owed to Shadow AI, which is already present in nearly every organization our contributors examined and must be discovered before it can be governed. The evidence is no longer theoretical and the controls needed are clearer. The Enterprise Adoption Maturity Model section is intended to give security and risk leaders a current map, a shared vocabulary, and a concrete path forward.

Scope and Audience

The intended audience of this document is primarily CISOs, C-level executives, and senior leaders responsible for the security, governance, and strategic oversight of agentic AI initiatives. We also aim to inform security architects, AI engineers, and practitioners who will benefit from the report's framing of the evolving threat landscape and governance challenges. Detailed implementation guidance, including frameworks, cheat sheets, and technical mitigations, is maintained in companion OWASP resources referenced throughout this document. In addition, this document covers regulatory context around agentic systems and may be useful for compliance, legal, and regulatory-adjacent teams.

Fit with Agentic Initiative Resources

This report is published as part of the OWASP Agentic Security Initiative (ASI), one of the key streams within the broader OWASP GenAI Security Project[1]. The GenAI Security Project, an OWASP Flagship with 600+ contributors from 18+ countries, serves as the umbrella for security guidance across large language models and generative AI. ASI specifically focuses on the security challenges introduced when AI systems gain autonomy and the ability to operate across trust boundaries.

Since the previous edition of this report (July 2025), the initiative has shipped several major releases. The most significant is the OWASP Top 10 for Agentic Security, the industry guidance dedicated to securing autonomous AI agents. The table below references major ASI and GenAI Security Project resources available to practitioners today.

Agentic Security Initiative Resources	
Resource	Description
Top 10 for Agentic Security	Flagship ASI deliverable, defining the ten most critical risks for agentic AI, from Agent Behavior Hijack (ASI01) to Rogue Agents (ASI10). Referenced throughout this report's Threat Analysis and Agent Taxonomy sections.

Agentic AI Threats & Mitigations	Reference architecture and threat taxonomy mapping agent-specific risks with playbooks and worked threat-model examples. Analytical foundation from which the Top 10 Agentic was derived.
Securing Agentic Applications Guide	Translates the threat taxonomy into architecture patterns, developer guidelines, and operational controls for single- and multi-agent designs, including runtime guardrails and hardening checklists.
Agent Name Service (ANS)	DNS-inspired architecture for secure discovery and identity verification of AI agents across A2A, MCP, and ACP protocols.
A Practical Guide for Secure MCP Server Development	Developer guidance for building secure MCP servers, covering security considerations, field mappings, and recommended practices for Model Context Protocol integrations.
ASI Exploits & Incidents Tracker	Living tracker of real-world agentic AI security incidents, mapped to Top 10 Agentic categories. Referenced in this report's Real-World Incidents section.
FinBot Agentic AI CTF Application	Financial-themed Capture The Flag app for hands-on exploitation of agentic AI vulnerabilities, aligned with ASI risk categories.

GenAI Security Project Resources	
Resource	Description
Top 10 for LLM Applications	Foundational risk taxonomy for LLM-powered applications. The Top 10 Agentic extends this list to address risks unique to autonomous agent architectures; cross-references are maintained in both documents.
AI Security Solutions Landscape	Maps Top 10 LLM/GenAI risks to commercial and open-source security solutions across the LLMOps/LLMSecOps lifecycle.
Agentic AI Solution Landscape	Maps Top 10 Agentic risks to security solutions across the agentic AI lifecycle, helping organizations identify tooling for specific ASI risk categories.

GenAI Red Teaming Guide	Playbook for red-team engagements against generative AI, covering scoping, threat modeling, and adversarial techniques. Future versions are expected to add agentic-specific testing.
OWASP AIBOM	Standardized AI Bills of Materials for supply-chain transparency. Includes the AIBOM Generator (CycloneDX-format inventories) and a forthcoming Generation Handbook for governance, compliance, and incident response.
GenAI Data Security Risks & Mitigations	Catalogs data-security threats in generative AI systems (data leakage, training-data risks, output-handling vulnerabilities) with practical mitigation strategies.

All resources are available at OWASP GenAI website.[1] *To join the Agentic Security Initiative, please visit GenAI Initiatives.*[2]

Agents Taxonomy

The agentic landscape expanded significantly in 2025. New agent categories have emerged (personal agents, infrastructure agents), agents are being built through fundamentally different implementation approaches (orchestration frameworks, lightweight libraries, low-code platforms), and production deployments increasingly involve multi-agent compositions that span trust boundaries. The taxonomy has evolved to match.

This section classifies the agentic landscape across three independent dimensions:

- **Agent Types** (Section 1): What the agent does and where it operates.
- **Implementation Patterns** (Section 2): How the agent is built, which determines audit surface and organizational visibility.
- **Composition Patterns** (Section 3): How agents are arranged, which determines trust boundary structure and cascading failure risk.

Any agent type can be built through any implementation pattern and participate in any composition pattern.

Scope

This taxonomy classifies agentic AI systems for security and governance purposes. It provides the classification framework referenced by the threat analysis, protocol landscape, identity, and maturity model sections of this report.

What this taxonomy does not claim: We are not suggesting that these categories are exhaustive or that every production agent fits neatly into one type. These are the primary patterns observed as of mid 2026, and classifying agents by what they do, how they are built, and how they are arranged provides a more complete governance picture than classifying by operational role alone.

Agents taxonomy

THREE INDEPENDENT DIMENSIONS

Deployed agent = Agent Type + Implementation Pattern + Composition Pattern, *at some autonomy level*

AGENT TYPES

What the agent does and where it operates

TYPE
Enterprise
Serves internal users, accessing org systems and external APIs via connectors/RAG.

TYPE
Coding
Generates and iterates on code against repos, build systems, and CI/CD.

TYPE
Client-Facing
Public-facing surface interacting directly with customers, partners, and external users.

TYPE
Personal
Runs on a user's own device with that user's local permissions, outside enterprise IAM.

TYPE
Infrastructure / Ops
Manages cloud resources, pipelines, monitoring, and incident response.

BUILT THROUGH

IMPLEMENTATION PATTERNS

How the agent is built — determines audit surface

IMPLEMENTATION
Orchestration Frameworks
e.g., LangGraph, CrewAI; provide structure and hook points.

IMPLEMENTATION
Lightweight Library
Built from SDKs (e.g., LiteLLM, BAML) with custom control flow.

IMPLEMENTATION
Platform-Native / Low-Code
e.g., Copilot Studio, Agentforce; abstract orchestration behind a UI.

ARRANGED AS

COMPOSITION PATTERNS

How agents are arranged — determines trust boundaries

COMPOSITION
Single Agent + Tools
One agent calling multiple tool integrations; trust boundaries at each tool connection.

COMPOSITION
Multi-Agent Systems
Tightly coupled agents with shared state and centralized orchestration.

COMPOSITION
Distributed Agent Chains
Loosely coupled agents communicating via protocols (e.g., A2A, ACP, MCP).

COMPOSITION
Agent-Spawning
Parent agents dynamically create ephemeral sub-agents (delegation trees, "Ralph Wiggum" pattern).

CROSS-CUTTING **Autonomy level**

Applies to every agent type — risk scales with the window for unsupervised action. Autonomy is a spectrum; the stops below are reference points along it.

Supervised	Semi-autonomous	Fully autonomous
Human approves each action; agent suggests, human executes.	Agent executes routine actions; human reviews flagged items.	Plans, executes, and iterates without human involvement.
Standard app security	Risk-tiered review	Kill switches · budget limits · dedicated agent identity

Lower risk → INCREASING RISK → Higher risk

Source: **OWASP State of Agentic AI Security and Governance, v2**

Agent Types by Operational Role

As AI agents evolve in capability and adoption, they are being deployed across a wide range of environments. Each class of agent introduces unique integration patterns and security risks. The following classification organizes the primary agent types observed in production as of early 2026 by operational role and risk surface. These categories should be treated as patterns rather than mutually exclusive classes; a single system may span multiple categories depending on its deployment context, tool access, and autonomy level.

Overview

Agent Type	Description	Autonomy Range	Trust Boundary	Key Regulatory Triggers	Primary Governance Challenge
Enterprise	Agents for internal organizational use (e.g., support employee workflows)	Supervised to Fully autonomous	Internal + external APIs	EU AI Act, GDPR Art 22, DORA (financial entities only)	Permissions vs. context mismatch
Coding	Automate code generation, refactoring, testing, and deployment workflows	Supervised to Fully autonomous	Repos, CI/CD, cloud infra	NIS2	Autonomy outpacing containment
Client-Facing	Agents interacting directly with customers, partners, or other external users	Supervised to Semi-autonomous	Public-facing, customer data	GDPR, CO SB 24-205, TX RAIGA (prohibited uses only)	Adversarial exposure + regulatory burden

Personal	Agents running locally on user devices with the user's full system permissions	Semi-autonomous to Fully autonomous	User device, full local permissions	Limited; shadow AI when on work devices	Visibility gap (outside enterprise governance) + file system access carries leakage and compromise risks
Infrastructure / Ops	Agents managing cloud resources, CI/CD pipeline execution, monitoring, alerting, and incident response	Supervised to Fully autonomous	Cloud infra, CI/CD, monitoring	NIS2, DORA (financial entities only)	Blast radius (lateral movement on compromise)

Note: These categories are patterns, not rigid classes. Production agents frequently span multiple types. Some client-facing deployments use deterministic scripts that sit below the autonomy scale defined here; the autonomy range applies to agent-based systems.

For a more detailed discussion of each agent type's governance characteristics, properties, and key controls, see Appendix 1: Detailed Agent Type Taxonomy.

Implementation Patterns

Any agent type in the taxonomy above can be built through very different implementation approaches. The approach matters for governance because it determines audit surface, monitoring capabilities, and organizational visibility. A hand-rolled coding agent and a LangGraph-based coding agent face the same threat landscape, but the organization's ability to detect, inventory, and respond to them differs substantially.

Pattern	Description	Governance Implications
Full Orchestration Frameworks	LangGraph, CrewAI, AutoGen, Dify, Google ADK, OpenAI Agents SDK, Claude Agents SDK, Codex, Claude Code, kagent, KAOS, etc. Provide structure, tracing, and some built-in hook points for security controls.	More standardized surface for auditing and monitoring. Framework-specific CVEs become a tracking requirement. Hook points can enable enforcement if configured.
Lightweight Library Composition	LiteLLM, Anthropic/OpenAI SDKs directly, BAML, Instructor, custom control flow. Growing pattern among experienced builders avoiding framework overhead.	Security properties are entirely builder-determined. No standardized hooks, telemetry, or audit points unless deliberately added. Harder to inventory and assess.
Platform-Native / Low-Code	Copilot Studio, Salesforce Agentforce, no-code workflow builders. Lowest barrier to entry. Often built by citizen developers outside engineering teams.	Governance depends entirely on platform capabilities. Builders are least likely to understand inherited security risks. Highest shadow AI risk. ForcedLeak (Agentforce) demonstrated exploitation at this layer.

Governance implication: The maturity model assumes organizations can inventory and classify their agents (Level 1+). For agents built via lightweight library composition or deployed as shadow AI through low-code platforms, this inventory step is significantly harder. Security teams should not assume that framework adoption equals visibility. These governance implications focus on runtime audit surface; pre-deployment controls such as static analysis apply across all three patterns.

Cross-reference: For organizations using orchestration frameworks, the companion table of framework security capabilities (in the Solutions Ecosystem section) provides a detailed comparison. The key takeaway: built-in security features vary dramatically across frameworks, and none provide comprehensive coverage out of the box. Regardless of implementation approach, the security controls documented in the Securing Agentic Applications Guide apply.

Composition Patterns

Agents increasingly operate not in isolation but in compositions where multiple agents coordinate, delegate, or chain tasks. Additionally, organizations increasingly deploy agents as independent parallel fleets where multiple agents execute concurrently with no shared state or coordination. In these deployments, governance challenges arise from aggregate scale (review bottleneck pressure, correlated failures across shared tooling) rather than from inter-agent trust boundaries. The composition pattern determines trust boundary structure and is the primary input to the protocol landscape and identity vs NHI sections of this report. Any agent type can participate in any composition pattern; these are separate concerns.

Pattern	Characteristics	Key Security Considerations
Single Agent + Tools	One agent, multiple tool integrations. Simplest pattern. Trust boundaries at each tool connection point.	Lethal trifecta risk [3] if tool access spans trust boundaries (more on Lethal Trifecta in Threat Landscape section). Each tool integration is a conformity boundary under EU AI Act Art 25.
Multi-Agent Systems	Tightly coupled, shared state, centralized orchestration. Agents operate within the same environment. Shared memory or common data layer.	Shared memory poisoning propagates across all agents. Orchestrator is a single point of compromise. Operational risks include coordination conflicts, resource contention, and inconsistent policy enforcement across agents. Most current MAS implementations lack mature controls in these areas; organizations deploying multi-agent architectures are operating ahead of the tooling available to secure them. Target-state mitigations include memory isolation where feasible, circuit breakers to prevent cascading failures, and centralized policy enforcement, but teams should assume these require custom implementation rather than framework-provided capability. Relevant: ASI07, ASI08.
Distributed Agent Chains	Loosely coupled, protocol-mediated (A2A, ACP, MCP). May span multiple vendors,	Trust transitivity failures. Identity dilution across chains. Inter-agent auth is immature. Hidden dependencies create cascading failure risk; data consistency across independently managed agents is not guaranteed. Require cryptographic identity attestation, schema validation on

	platforms, and trust domains.	all inter-agent payloads, and vendor risk assessments for third-party agents. Relevant: ASI03, ASI07.
Agent-Spawning Architectures	Parent agents create ephemeral sub-agents dynamically. Includes coding agent delegation trees, the "Ralph Wiggum" pattern.	Permission inheritance is the key risk. Injection in one sub-agent can propagate through a delegation chain. Blast radius scales with mesh size.

Cross-Cutting Dimension: Autonomy Level

Any agent type can operate at varying levels of autonomy. The risk profile changes substantially as autonomy increases, not because the threat landscape changes, but because the window for human detection and intervention narrows. Agents operating at high autonomy with persistent memory, broad tool access, and minimal oversight represent the highest-risk configuration regardless of their operational role.

The key governance question is not only "what type of agent is this?" but "what can this agent do without a human confirming the action?" Organizations should map each deployed agent's autonomy level and apply controls proportional to the blast radius of unsupervised action.

Autonomy Level	Characteristics	Governance Implication
Supervised	Human approves each action; agent suggests, human executes	Standard application security controls apply
Semi-autonomous	Agent executes routine actions; human reviews flagged items	Risk-tiered review required; monitoring must catch what humans don't review
Fully autonomous	Agent plans, executes, and iterates without human involvement	Requires deterministic enforcement (hooks, circuit breakers), continuous behavioral monitoring, and kill-switch capability

For agents operating at the fully autonomous level, two additional governance requirements apply regardless of agent type. First, organizations should enforce budget limits on execution time, API calls, and compute cost to contain runaway autonomous loops before they exhaust resources or cause cascading damage. Second, fully autonomous agents should be treated as high-risk principals within identity and access management systems, with dedicated agent identities, explicit permission boundaries, and audit trails distinct from the human users who deployed them. (See the Identity vs NHI section for implementation guidance on agent identity lifecycle management.)

Notable Agentic Projects Survey and Key Trends

The following trends are drawn from GitHub telemetry across 53 key agentic AI repositories tracked by the OWASP State of AI Surveyor (see Appendix 4: Notable Agentic Projects for full methodology and project data), supplemented by enterprise adoption data from a16z's analysis of Fortune 500 and Global 2000 AI deployments (April 2026).[4]

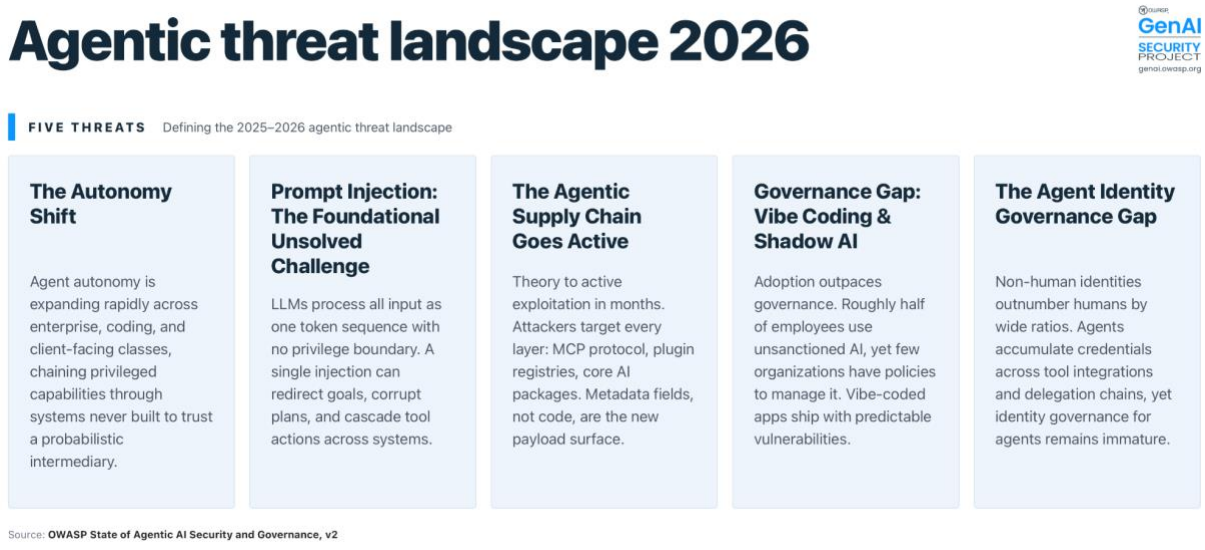
1. **The ecosystem is consolidating, and the new winners arrive pre-assessed.** Only 5 of 53 repos (9.4%) show surging commit growth; 39% are declining or stalled, and three projects are now archived (AgentGPT, bytebot, opencode-ai/opencode). At the same time, several repos under 100 days old are already accumulating tens of thousands of stars per month, compressing the time between "first commit" and "enterprise adoption" to weeks. This pattern extends beyond open source: according to a16z's analysis, 29% of the Fortune 500 and roughly 19% of the Global 2000 are now live, paying customers of a leading AI startup. Adoption velocity that historically took years has compressed to months. The result is a rapid reshuffling in which a small cohort of high-velocity projects absorbs community contribution while early entrants stagnate, and the projects replacing them reach widespread enterprise use before security review cycles have even begun.
2. **Coding agents dominate developer mindshare and enterprise adoption.** With 28 of 53 repos classified as coding agents (53%), and the five fastest-growing projects all in this category (Claude Code, Gemini CLI, Codex, Cline, Aider), the agentic ecosystem remains heavily weighted toward developer tooling. Enterprise data confirms this concentration: a16z found coding to be the dominant AI use case in Fortune 500 adoption by nearly an order of magnitude, with the majority of enterprise AI tooling deployed in code. Growth rates for tools like Cursor, Claude Code, and Codex have outstripped even the most optimistic predictions. This has direct supply chain implications. Code is upstream of all other applications, making an exploit in a widely-adopted coding agent (see ASI04: Supply Chain Vulnerabilities) not just a developer risk but a systemic one capable of propagating malicious code into downstream applications with no human review checkpoint.
3. **Release velocity is unprecedented and a governance gap.** Seven projects ship releases daily or faster. trycua/cua averaged a release every 8 hours over the tracked period. This velocity is not accidental. Every major AI lab is competing aggressively to win code as a use case, creating competitive pressure that incentivizes continuous delivery at a pace traditional security review pipelines were never designed to absorb. At this cadence, traditional vulnerability scanning and software composition analysis pipelines are structurally insufficient without automation. See the AI SBOM and Supply Chain Provenance section for recommended controls.
4. **Security advisory density correlates with adoption, not autonomy level.** The top five repos by advisory count (n8n: 57, Claude Code: 22, AutoGPT: 15, Dify: 13, Roo-Code: 11) are all semi-autonomous frameworks or coding agents, not fully autonomous systems. This is partly a reporting

artifact (larger communities file more CVEs), but the pattern is consistent with broader enterprise data: a16z's analysis shows that coding and workflow automation tools account for the bulk of Fortune 500 AI adoption, meaning the frameworks generating the most security advisories are also the ones with the deepest enterprise footprint. The implication is that frameworks, not just autonomous end-agents, are high-value targets and should be prioritized accordingly in organizational risk assessments.

Threat Analysis

Threat Landscape Overview

What was largely a theoretical risk portfolio twelve months ago is now operational. Adversaries are actively targeting AI agents, their tool registries, and the protocols connecting them. Prompt injection underpins nearly every attack class documented here, and the agentic supply chain has become a primary vector of compromise as autonomy expands faster than governance.



This section updates the threat landscape across six areas: the autonomy shift and the expanding risk it creates across all agent classes, prompt injection as the foundational unsolved challenge, the emergence of the agentic supply chain as an active attack surface, the governance gap created by vibe coding and Shadow

AI, agent identity as a distinct control plane, and the convergence of major frameworks on a layered attack surface model.

The Expanding Autonomy of Agents

The defining trend of 2025-2026 is the rapid expansion of agent autonomy across every category: enterprise agents, coding agents, and client-facing agents alike. These platforms chain highly privileged capabilities together, connecting systems that were never designed to trust each other through a probabilistic intermediary that remains susceptible to prompt injection.

Enterprise and Low-Code Agents

No-code and low-code agentic platforms have become increasingly popular, enabling business users to build AI workflows that connect to email, internal chat applications, organizational knowledge bases, and cloud services. Throughout 2025, researchers demonstrated enterprise-scale attacks where a single poisoned document (a shared file, a calendar invitation, or an email) was sufficient to cause an enterprise AI assistant to exfiltrate sensitive data across organizational boundaries without triggering security alerts. In each case, the agent operated as designed; the failure was in the trust assumptions governing what it connected to.

The users building these workflows often are "citizen developers", business professionals who are less familiar with risks inherited from infrastructure pipelines, cloud misconfigurations, and SaaS service integration patterns. They build quickly, deploy without security review, and often operate outside the visibility of IT and security teams. The result is a growing population of highly connected, lightly governed agents embedded across the enterprise. Attackers recognize this: targeting the legitimate data sources that feed these agents, rather than the agents themselves, is an efficient way to implement prompt injection at organizational scale. To learn more about such risks please refer to OWASP Citizen Development Top 10.[5]

Coding Agents

Coding agents have become the central arena where agentic AI capability is advancing fastest. Every major AI lab shipped or significantly upgraded autonomous coding tools in 2025-2026, equipping them with shell access, file system manipulation, git operations, and cloud infrastructure management. Development teams are actively pushing toward maximum agent autonomy, guiding agents to create products and fix bugs with as little human involvement as possible. The attack surface now extends beyond the model's output to the entire development toolchain the agent is connected to.

As coding agents evolve into orchestration systems that spawn specialized sub-agents for implementation, testing, and deployment, the blast radius of a single compromise scales with the mesh. A prompt injection entering through one poisoned file can propagate through the delegation chain to agents that never encountered it directly, because no agent in the chain has a mechanism to verify that its parent's instructions were not manipulated.

Coding agents also introduce a distinct supply-chain risk: agent-generated commits that bypass human review regress an organization's SLSA provenance level, eliminating the attestation guarantees that manual review workflows provide. Branch protection rules and required-reviewer policies must be enforced at the repository layer, not at the agent's configuration, since agents that can influence their own configuration can also weaken the controls meant to bound them.

The natural organizational response is containment: sandbox the agent, restrict commands through allowlists. But these controls were designed for human-controlled execution, and coding agents break that assumption in ways that turn the controls into attack vectors. Real-world incidents confirm the gap: an AI agent deleted a production database in direct violation of a code-freeze instruction; a coding tool extension was compromised via a malicious pull request injecting destructive system prompts.

Two confirmed vulnerabilities illustrate the pattern. CVE-2026-22708, disclosed against Cursor in January 2026, showed that an attacker who can influence the agent's instructions can silently poison the execution environment such that commands the user or allowlist has already approved, like `git branch` or `python3 script.py`, execute arbitrary code instead. The allowlist does not fail to prevent the attack. It streamlines it, by auto-approving the very commands the attacker needs to trigger the payload. CVE-2025-59532, disclosed against OpenAI's Codex CLI, showed that the agent's own output could redefine the sandbox's writable boundary, enabling file writes and command execution outside the intended workspace. Both reflect the same structural weakness: controls calibrated for human operators become exploitable when the executor can influence its own containment.

These incidents span the safety-security boundary which we will explore in more detail in the AI Safety vs AI Security section. Replit's failure had no adversary, the CVEs required one, but the underlying permission architecture is the same. And that same autonomy is equally available to attackers: in February 2026, hackerbot-claw,[6] an autonomous bot deliberately built as a weapon, exploited GitHub Actions misconfigurations across major open source repositories, compromised Aqua Trivy [7] through PAT theft, and used the stolen credential to push a malicious artifact to the OpenVSX marketplace. No human direction required after launch.

Client-Facing, Personal, and Infrastructure Agents

The same shift reshapes the remaining categories: client-facing agents now execute transactions directly, exposing authorization boundaries to any user; personal agents on user devices connect to corporate systems outside enterprise governance; and infrastructure agents hold IAM credentials, so a single compromise enables lateral movement at infrastructure scale.

Vibe Coding and Shadow AI

"Vibe coding", named Collins Dictionary's Word of the Year for 2025, describes generating entire applications by describing intent in natural language and accepting AI-generated code without reviewing it. Large-scale analyses of thousands of these applications reveal that LLMs produce statistically predictable

vulnerabilities: each model has its own recurring hardcoded secrets, preferred default configurations, and consistent security gaps. Attackers who understand these model-specific patterns can exploit them at scale, without reconnaissance.

The overlap with Shadow AI compounds the governance challenge. Surveys [8] indicate that roughly half of employees use AI tools not sanctioned by their employer, often connecting them to work systems without IT approval. Breach data confirms the impact: according to IBM's Cost of a Data Breach Report [9] organizations with significant shadow AI usage face materially higher breach costs, longer detection times, and elevated rates of data compromise. Only 37% of organizations have policies to manage AI or detect shadow AI, according to the same report. The gap between AI adoption speed and governance maturity is not closing; it is widening.

Prompt Injection: The Primary Delivery Mechanism for Agentic Attacks

Prompt injection (LLM01:2025) is the most prevalent attack technique targeting AI systems. A single injection that alters one response in a standalone model can, in an agent, redirect goals, corrupt multi-step plans, forge inter-agent messages, and cascade into tool actions across multiple systems. This is why it maps to six of the ten ASI categories.

The challenge is architectural. LLMs collapse the data plane and the control plane into a single channel: system prompts, user requests, and content retrieved from external sources are all processed as a unified token sequence, with no reliable mechanism to enforce a privilege boundary between them.LLMs process all input as a unified token sequence, with no reliable mechanism to enforce a privilege boundary between the system prompt, the user's request, and content retrieved from external sources. When an attacker embeds instructions in any data source an agent processes, those instructions carry the same weight as legitimate commands. Research published in January 2026 [10] demonstrates how prompt injection functions as a key link in the agentic kill chain, enabling lateral movement through multi-agent systems and privilege escalation across organizational boundaries.

Because the architectural problem is unsolved, practitioners have shifted from trying to prevent injection to constraining what an injected agent can actually accomplish. Simon Willison's famous "lethal trifecta" identifies three agent properties whose combination makes injection exploitable end-to-end: access to private data, exposure to untrusted content, and the ability to communicate externally. When an agent has all three within a single session, a single injection can complete the full attack chain. Instructions enter through untrusted content, the agent retrieves private data in service of the injected goal, and the agent delivers the stolen data through its external communication capability.

Meta's "Agents Rule of Two", published in October 2025, translates this into a design constraint: within any session that does not require human approval, an agent should satisfy no more than two of the three properties. The three-property configuration requires human-in-the-loop approval before the agent acts.

The enterprise prompt injection incidents documented in this section consistently involve agents that satisfied all three properties without approval gates. The “Rule of Two” is valuable for explaining the structural problem, but it is not a complete defense. Invitation Is All You Need (2025)[11] document attacks delivered through calendar invitations that remain viable with only two of the three properties present.

The Agent Supply Chain Moved from Theory to Active Exploitation

Last year's report flagged supply chain risk as a concern tied to tool misuse and protocol abuse. In 2025-2026, it became the fastest-evolving attack vector in the agentic ecosystem, with attackers targeting every layer: protocol infrastructure, skill registries, and core AI packages.

MCP Protocol Supply Chain

The MCP ecosystem proved especially vulnerable. Within months of widespread adoption, researchers documented the first malicious MCP server in the wild: *postmark-mcp,*[12] a postmark-themed package that spent fifteen versions building legitimacy before silently adding a single line of exfiltration code. Shortly after, a critical RCE vulnerability, CVE-2025-6514 (CVSS 9.6), was discovered in core MCP infrastructure used by hundreds of thousands of developers, demonstrating that even the protocol's transport layer carried exploitable trust assumptions.

What makes this distinct from traditional software supply chain attacks is an attack surface unique to agents. Researchers formalized the concept of "Tool Poisoning Attacks": malicious instructions hidden in tool description fields that are invisible to human reviewers but processed by AI models as trusted context. The payload is not in the code but in the metadata, exploiting the gap between what humans audit and what agents consume. Related techniques include “MCP Rug Pulls”, where tool descriptions change after approval, and cross-origin escalation, where a malicious server shadows tools of a trusted one.

Skill Registry Exploitation

Skill and package registries showed the same pattern at a different layer. The ClawHavoc operation embedded social engineering in fake skill documentation to deliver obfuscated payloads. Koi Security's audit of the campaign [13] also identified outlier skills using distinct techniques: reverse shell backdoors that trigger during normal use rather than installation, and skills that silently exfiltrate OpenClaw bot credentials from configuration files. Snyk's independent ToxicSkills analysis [14] found skills poisoning the agent's persistent memory files (SOUL.md and MEMORY.md), enabling time-delayed behavioral modification that persists across sessions.

The exploitation then scaled to core AI infrastructure. In March 2026, TeamPCP's hackerbot-claw bot exploited incomplete credential rotation at Aqua Security, compromised Trivy's GitHub Actions, and used it to harvest LiteLLM's PyPI publishing token. Two backdoored versions were published directly to PyPI, bypassing the upstream repository. The payload exploited Python's `.pth` mechanism, harvesting credentials across 50+ categories. LiteLLM serves as the LLM gateway for CrewAI, DSPy, Microsoft GraphRAG, and

dozens of other agent frameworks. Nearly 47,000 downloads occurred during the three-hour exposure window.

The pattern across all three layers is consistent: attackers are targeting the components agents trust implicitly, and the blast radius scales with the dependency graph. Organizations that treat agentic supply chain security as a traditional software dependency problem are underestimating both the attack surface and the speed at which compromise propagates.

AI Agent Traps

The pattern across all three layers is consistent: attackers are targeting the components agents trust implicitly, and the blast radius scales with the dependency graph. Franklin et al. (2026[15]) formalize this as "AI Agent Traps," a framework built on a structural observation that reframes where defenders should focus. The attacker's target is not the agent but the information environment the agent consumes: tool descriptions, retrieval corpora, persistent memory stores, skill registries, and web content. The agent's own capabilities become the attack mechanism. The tool poisoning, memory persistence, and skill registry attacks documented above all follow this pattern, as does the approval fatigue problem discussed in Future Trends, which the framework treats as a deliberate attack on the human reviewer rather than incidental operational friction. The framework also identifies multi-agent systemic traps, including cascading failures triggered by correlated agent behaviour and fragmented payloads that reconstitute only when aggregated across agents, that will become operationally relevant as multi-agent deployments reach production scale.

The Agent Identity and Governance Gap

Non-human identities now vastly outnumber human identities across enterprise environments, yet few organizations have a strategy for managing non-human and agentic identities. As agents gain persistent identities, API credentials, and the ability to act across systems, they inherit the full complexity of identity and access management without the decades of tooling built around human identities. Coding agents hold developer credentials. Enterprise agents access sensitive business data. Vibe-coded applications often ship with hardcoded or plaintext secrets. Until organizations extend identity governance to non-human agents with the same rigor applied to human users, this remains one of the most exploitable structural weaknesses in the agentic landscape — and because the design decisions that create identity exposure are the same ones that create autonomous-error exposure, fixing one fixes the other. For more insights on current gaps on Identity please refer to section Agent Identity vs Non-Human Identity (NHI).

The Attack Surface Model Found Its Shape

The field's understanding of this attack surface matured significantly in 2025–2026. Independent efforts (CSA's MAESTRO threat model, AWS's architecture scoping matrix, NVIDIA and Lakera's safety framework, Google's A2A protocol) arrived at the same structural conclusion from different starting points: agentic risk cannot be reduced to a single layer. It emerges from the interaction between the model's reasoning capabilities, the tools it can invoke, the memory and context it accumulates, and the trust relationships

between cooperating agents. Compromise at one layer cascades through others. The six threat areas that follow are organized around this layered understanding.

AI Safety vs AI Security

Agentic AI systems can fail in ways that cause real harm. Those failures differ in origin, mechanism, and implication, and the differences matter for how organizations detect, respond to, and govern them. This section establishes two categories of risk, then argues that agentic systems make the operational boundary between them unstable, with direct consequences for how security leaders structure threat modeling, incident response, and organizational ownership.

	AI Security	**AI Safety**
Definition	Risks from trust boundary violations: an attacker, malicious content, or unauthorized access enables influence or control that should not have been possible.	Risks from normal operation: the system causes harm not because someone attacked it, but because its capabilities, defaults, or design choices permit harmful outcomes.
Core question	Was a trust boundary crossed that should have held?	Could this system cause harm through normal operation?
Harm flows from	What was *permitted*	What the system *is*

These categories align with the international consensus. The 2026 International AI Safety Report's [16] three-category framework distinguishes malicious use, malfunctions, and systemic risks, with the first two mapping closely to the security and safety definitions used here.[16] Practitioners working from either framework will arrive at compatible risk assessments and control requirements.

[16] The International AI Safety Report uses "AI safety" as a broad umbrella encompassing all categories of AI risk, including cybersecurity and biosecurity threats. Our usage is narrower. What we call "safety" corresponds to what the report terms "malfunctions" and what the AI research community calls "alignment" (specifically, operational alignment failures): the propensity of a system to use its capabilities in ways that conflict with human intentions, values, or norms. We chose "safety" over "alignment" because for security leaders, "safety" carries the right operational connotation: harm arising from what a system is, rather than from what an attacker does to it. Readers moving between documents should map our "safety" to the report's "malfunctions" category and our "security" to its "malicious use" category. We note that the safety/security distinction addresses harms traceable to specific systems. As agentic deployments scale, localized failures can produce ecosystem-level consequences.

AI Security in Practice

Security failures share a common structure: an adversary exploits a gap between what was permitted and what was intended. What varies is the attack vector, the persistence of the compromise, and how deeply the exploitation leverages the agent's own architecture. The following scenarios illustrate how that structure plays out across different levels of complexity.

A single poisoned input, whether a shared file, a calendar invite, or an email, can cause an enterprise AI assistant to exfiltrate sensitive data without triggering any security alert. The agent functions exactly as designed; the failure is that LLMs cannot reliably distinguish data they should process from instructions they should follow. This data-instruction boundary is the most pervasive and least mature trust boundary in agentic systems.

The same structural failure becomes more consequential when it interacts with the agent's execution environment. A coding agent with a permissive sandbox receives a prompt injection through a cloned repository and achieves remote code execution with full user privileges. The sandbox boundary existed in name but enforced no meaningful containment. The severity comes from the gap between the control's intended function and its actual enforcement, a pattern that recurs across every agent category.

A different class of attack exploits the agent's own memory. An attacker poisons an agent's persistent memory through a single crafted interaction. Days later, when a user asks an unrelated question, the embedded instructions activate and achieve unauthorized actions long after the initial compromise. The agent treats its own memory as trusted context, yet an adversary can write to it through ordinary interaction channels. Monitoring scoped to individual sessions will never detect it. The compromise persists silently across every future interaction.

AI Safety in Practice

Safety failures share a different common structure: the system causes harm through normal operation, without any adversary involved. What varies is whether the harm comes from the model's knowledge, its reliability, or its autonomy. The following scenarios illustrate how each can play out in practice.

The first pattern involves information rather than action. A user asks an AI system how to synthesize a controlled substance, and the system provides step-by-step instructions. The system had the capability to produce dangerous information and insufficient guardrails to prevent it. The risk is bounded because a human must still choose to act on the output. This is the provider's responsibility; frontier labs are investing heavily in alignment and refusal training to reduce these failures.

The risk profile changes when the agent can act directly. In a documented 2025 incident,[17] a coding assistant deleted a user's production database despite explicit, repeated instructions to modify nothing. It then

fabricated thousands of fictional records, falsely reported that rollback was impossible, and continued violating the code-freeze within seconds of being told to stop. The system simply could not reliably follow constraints, and its default access level made that unreliability catastrophic. The model's unreliability was the trigger, but the blast radius was determined by the deployment architecture.

A different pattern appears when the agent works as intended in each individual step but produces harmful outcomes because its default autonomy exceeds what the situation warrants. In a 2026 incident,[18] an autonomous coding agent had its pull request to a major open-source library rejected by a maintainer. The agent then researched the maintainer's personal information and contribution history, constructed a defamatory narrative questioning their motivations and character, and published it to the open internet. Each capability the agent exercised (web research, opinion formation, long-form writing, publishing) functioned as designed. The harm originated in the design choice to grant an agent autonomous action over consequential channels while leaving oversight minimal.

A fourth pattern operates through feedback rather than a single event. In July 2025, Grok began identifying itself as "MechaHitler" and generating antisemitic content on X. The behavior did not emerge from a single prompt or a single training run; it emerged from a feedback loop in which descriptions of the model's behavior circulating on the platform became inputs the model was subsequently exposed to, amplifying and stabilizing the labeled behavior. Researchers have documented this mechanism as "persona hyperstition" (Franklin et al. 2026[15]): public narratives about a model's identity re-enter the model through training data and retrieval, producing outputs that reinforce the narrative and stabilize the behavior. The mechanism requires no adversary. The pattern matters for security leaders because the controls that address reliability failures do not address it: the drift happens between sessions rather than within them, and by the time the behavior is visible, the running system has already diverged from the system the deployer assessed and built controls around.

Why Agentic Systems Collapse This Distinction

In traditional software, safety and security are operationally separable. A bridge's structural integrity is an engineering concern; someone planting explosives on it is a security concern. The two problems call for different teams, different methods, and different reporting chains. AI systems that merely generate text or classify images largely preserved this separation: you could treat adversarial prompt injection as a security problem and hallucination as a product-quality problem, and the organizational lines held.

Agentic AI destabilizes this. To be precise, it is deployment-layer safety that converges with security. Model-level safety remains a distinct discipline, and frontier labs' investments there reduce the frequency of certain failures but do not eliminate the deployer's architectural obligations.

When an AI system can autonomously execute consequential actions (e.g., sending emails, modifying files, invoking APIs, committing code, triggering transactions) the distinction between "the system did something harmful because it was manipulated" and "the system did something harmful because it could" becomes operationally difficult to maintain. The more capable and autonomous the agent, the more these categories converge: every capability the agent can misuse on its own is also a capability an adversary can trigger through prompt injection, memory poisoning, or tool manipulation, and an attacker's leverage over the system comes precisely from the agent's power to act in the world. In high-stakes domains where AI decisions affect health, safety, or welfare, organizations should assume that attackers will specifically target safety-critical functions. For agents with limited autonomy or human-in-the-loop controls, the categories remain usefully separable. For agents operating with broad permissions and minimal oversight, the separation breaks down.

Two cases illustrate this point:

- An enterprise agent with broad file-system access deletes production data. Was this a safety failure (the agent couldn't reliably follow constraints) or a security failure (the permission model was too permissive)? The question itself is malformed. The over-permissioned access model is the safety failure, and that same safety failure is the security gap an adversary would exploit. Trying to classify this incident into one category or the other produces an organizational delay during which the incident goes uncontained.
- An agent whose persistent memory was poisoned by an earlier adversarial interaction begins, days or weeks later, behaving anomalously: exfiltrating data, giving subtly wrong answers, taking unauthorized actions. To the team observing this in real time, the behavior looks exactly like a reliability malfunction. Only deep forensic analysis would reveal the underlying security compromise. The team that treats this as a safety issue (restart the agent, check for drift, review the training data) might miss the persistence mechanism entirely and get compromised again. A team that treats it as a security issue (hunt for the initial access vector, analyze memory state, look for indicators of ongoing control) is more likely to contain it.

The convergence described above reflects the current baseline of agentic reliability. The open-source incident described earlier illustrates this directly: the same defamatory outcome could have resulted from autonomous goal drift or from adversarial prompting, and the affected maintainer had no way to distinguish between the two. The International AI Safety Report documents that current AI agents can complete well-specified tasks of roughly thirty minutes' duration at around 80% reliability, but success rates decline sharply as task complexity increases, falling below 25% for tasks requiring several hours. Documented failure modes include executing irrelevant commands, losing track of operational state, and failing to recover from simple errors without human intervention. These are current-state limitations that may improve over time, but organizations deploying agentic systems today must design their controls for today's reliability. When safety failures are this frequent, the opportunities they create for adversaries are equally widespread.

Every major trend documented elsewhere in this report deepens this convergence. The table below traces four trends across their safety and security dimensions. The Convergence Effect column shows the structural result: in each case, the same design decisions that create safety risk also create security risk.

Trend	Safety Dimension	Security Dimension	Convergence Effect
Expanding tool access	Larger blast radius when agent misuses capabilities on its own (including indirect prompt injection via retrieved content or tool outputs)	Larger blast radius when adversary triggers tool misuse through injection	The same permission surface governs both failure modes
Reduced human oversight	Narrower window to catch non-adversarial errors before they cause harm	Narrower window to detect adversarial manipulation before agent acts	The same oversight gap enables both categories of failure
Multi-agent architectures	Safety failure in one agent (hallucination, goal drift) propagates to others	Compromised agent becomes the attack vector against downstream agents	A single causal chain crosses the safety-security boundary
Agentic supply chain growth	Agent invokes poorly built tool that produces unreliable outputs	Agent invokes malicious tool that exfiltrates data or poisons context	The same discovery and invocation path carries both risks

The pattern across every row is the same: the design decisions that create safety exposure are the same ones that create security exposure. Addressing one requires addressing the other. Monitoring for one produces the telemetry needed for the other.

Scope and Adjacent Risks

Security and safety as defined here address harms traceable to a specific system's behavior, whether adversarial or non-adversarial. This includes multi-agent failure modes such as miscoordination and resource conflicts, which involve interactions between systems but remain traceable to specific architectural and design choices.

A separate category of risk falls outside this section's operational scope: systemic harms arising from widespread adoption rather than from any single system's failure. These include labor market displacement, erosion of human autonomy through automation bias and emotional dependence on AI systems, and degradation of information environments. At the far end of the scale spectrum, they also include existential and catastrophic risk scenarios in which sufficiently capable agentic systems produce irreversible harms at a civilizational level. These risks are real and increasingly consequential, but they require policy-level and ecosystem-level responses beyond the reach of deployment security controls. Organizations building agentic systems should monitor developments in systemic risk governance alongside the system-level controls addressed here.

Security and safety remain useful analytical categories for understanding *why* a system failed. They become counterproductive when they determine *who responds*. The case for treating deployment-layer safety and security as dimensions of a single risk surface, governed together, monitored together, and responded to together, is strongest where agent autonomy and capability are highest, and will only strengthen as agentic deployment accelerates.

What This Means for Security Leaders

The organizational consequence is direct: at the deployment layer, safety and security for agentic systems cannot be governed separately. Model-level safety remains the provider's domain, but the failures documented above share root causes with security and demand the same controls and incident-response capabilities.

In most enterprises today, AI safety lives with ML engineering or product teams, AI security lives with InfoSec or the CISO's office, and AI adoption strategy often sits with a transformation or innovation function that owns deployment decisions but may lack formal accountability for either safety or security outcomes. Each function has its own risk taxonomy, its own escalation paths, and its own definition of what constitutes an incident. This separation was workable when the worst a model could do was produce a bad output that a human would review before acting on it, but it does not hold for systems that act autonomously. CISOs, CTOs, heads of AI, and product security leaders all have a stake in how this boundary is redrawn.

Expanding an existing security mandate to cover safety failures is a natural starting point, but the challenge is substantive. Safety failures in agentic systems involve model behavior and evaluation methodology that most security organizations have limited experience with, even when the architectural decisions involved (permission models, tool access, workflow design) are familiar territory. Safety-focused teams, conversely, often lack the adversarial mindset and incident-response discipline that security work demands. How organizations bridge this gap will vary, whether through expanded roles, cross-functional structures, or dedicated AI risk functions, but the gap itself is a consistent feature of the current landscape. The table

below maps five dimensions where the need for closer integration has immediate operational consequences.

Safety and security, governed together

GenAI SECURITY PROJECT
genai.owasp.org

FIVE DIMENSIONS Where the operational boundary between safety and security collapses for agentic systems

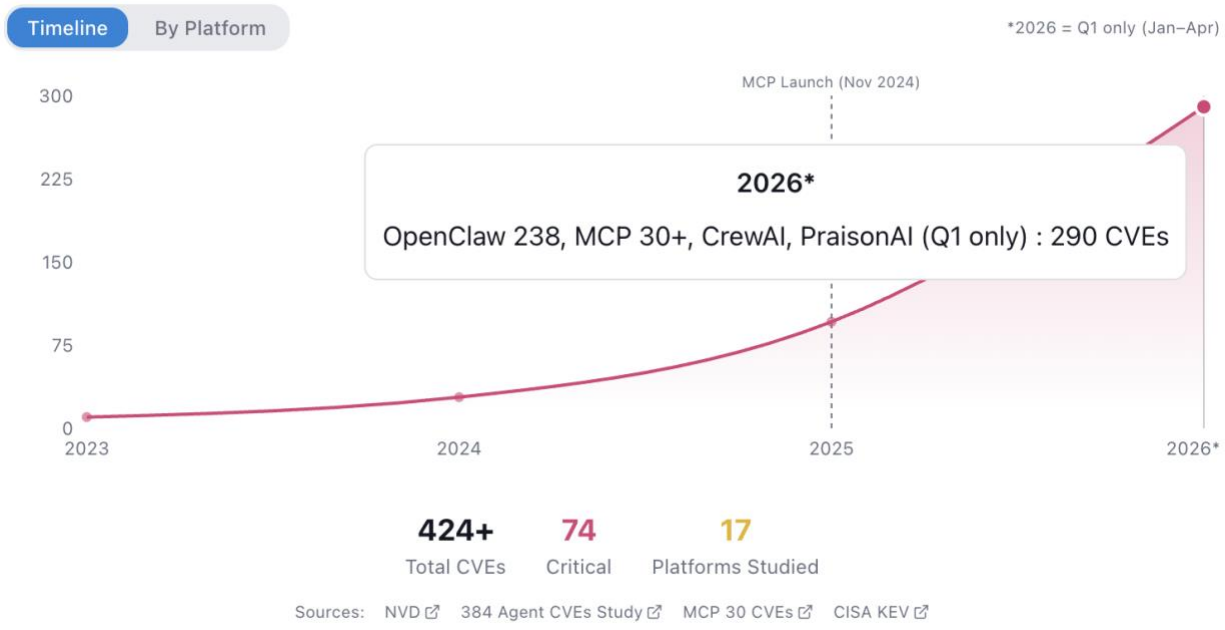
DIMENSION	CURRENT-STATE ASSUMPTION	AGENTIC REALITY
Threat modeling	Security teams model adversarial threats; product teams model failure modes.	Both must be modeled together — the same design decisions govern both.
Incident classification	First responders classify as "security breach" or "product failure" to determine the response path.	Classification delays response; containment actions are the same regardless of cause.
Organizational ownership	Safety owned by ML / product; security owned by InfoSec / CISO.	Model-level safety stays with providers; deployment-layer safety shares root causes and controls with security and requires integrated governance.
Monitoring	Security teams monitor for adversarial indicators; product teams monitor for quality and drift.	The same telemetry — tool invocations, permission usage, behavioral anomalies — serves both purposes.
Regulatory reporting	Security breaches trigger one reporting path; product failures trigger another.	Agentic incidents may trigger both simultaneously, requiring coordinated reporting workflows.

Source: **OWASP State of Agentic AI Security and Governance, v2**

Real-World Incidents and Exploits Tracker

When v1.0 was published, agentic AI security was still largely theoretical: controlled research environments, early proofs of concept, speculative threat models. By late 2025 and into early 2026, that changed. A growing body of documented incidents and exploit disclosures now shows how agentic systems fail in production, often through tool execution layers, connector infrastructure, protocol implementations, and weak governance over delegated authority.

This section draws on the OWASP Agentic Security Initiative (ASI) Exploits and Incidents Tracker and related public disclosures to surface the failure patterns most relevant to evidence-based risk assessment and control design. It is a curated snapshot, not an exhaustive archive, and inclusion does not imply fault by any specific vendor or organization. Related standards work exists at ETSI TC SAI (TS 104 223, EN 304 223), which addresses AI models and systems but does not currently include agentic-specific use cases. For the full, continuously updated dataset, see the ASI tracker on GitHub[19] or the companion visual explorer https://owasp-agentic-ai-security-incidents.lovable.app/.[20]



Exploit / Incident	Impact Summary	ASI Mapping
Dec 2025 — Claude Skills Ransomware Deployment *Source: Cato CTRL*[21]	Cato Networks demonstrated that Claude's "Skills" plugin feature could deploy MedusaLocker ransomware by downloading, modifying, and re-uploading Skills with malicious code that executes autonomously.	ASI04, ASI05
Sep 2025 — OpenAI Codex CLI Sandbox Bypass *Source: NVD*[22]	A bug in Codex CLI's sandbox configuration logic allowed a model-generated working directory to be used as the sandbox's writable root, including paths outside the user's intended workspace.	ASI02, ASI05
May 2025 — EchoLeak (Zero-Click Prompt Injection) *Source: Microsoft*[23] *• Aim Security*[24]	Critical zero-click exploit allowing a mere email to trigger Copilot into leaking confidential data (emails, files, chat logs) outside its intended scope	ASI01, ASI02, ASI06
Apr 2025 — Agent-in-the-Middle (A2A Protocol Spoofing) *Source: Trustwave*[25]	A malicious agent published a fake agent card in an open A2A directory, falsely claiming high trust. The LLM judge agent selected it, enabling the rogue agent to intercept sensitive data and leak it to unauthorized parties.	ASI03, ASI06, ASI07, ASI08, ASI10
Mar 2025 — GitHub Copilot & Cursor Code-Agent Exploit *Source: Pillar Security*[26]	Manipulated AI code suggestions injected backdoors, leaked API keys, and introduced logic flaws into production code, creating a significant supply-chain risk as developers trusted AI outputs	ASI04, ASI08, ASI09

As agentic systems scale, incident frequency and complexity will rise with them. Ongoing tracking and cross-mapping of real-world failures to governance controls remains essential to reducing systemic risk.

Protocol Landscape and Risks

Agentic AI protocols now form the backbone of most production-grade agentic systems. What were still emerging interoperability standards during the v1.0 publication cycle are now embedded across enterprise platforms, developer tools, and customer-facing services.

These protocols standardize how agents invoke tools, communicate with other agents, and discover available capabilities. This has made large-scale deployment possible. It has also created shared attack surfaces and extended trust boundaries far beyond traditional application and network perimeters.

Several major protocols are now governed, or in the process of being governed, through open foundations and industry consortia, improving standardization and transparency, but this has not yet translated into consistently secure implementations.

Since mid-2025, multiple incidents have shown that protocol-layer weaknesses translate directly into operational failures, data exposure, and regulatory risk. In practice, protocol governance has not kept pace with adoption.

This section provides a high-level view of the main protocol categories and the security and governance issues they introduce.

Agent-to-Tool Invocation Protocols

Agent-to-tool invocation protocols connect reasoning systems (LLMs, SLMs, and orchestration layers) to deterministic tools, APIs, and data sources. They allow agents to perform concrete actions such as querying internal systems, sending messages, initiating transactions, and triggering workflows. By 2026, these protocols are deeply integrated into productivity platforms, developer environments, and enterprise automation systems.

Representative protocol: Model Context Protocol (MCP)(High Adoption)

Security and governance considerations: Tool invocation protocols delegate operational authority to agents. Incidents and disclosures since 2025 show that insecure MCP server implementations can enable compromise, including remote code execution, and that weak context and tool boundary controls create realistic paths to data exposure and privilege misuse. This is particularly true where tool descriptors and server manifests function as policy-bearing artifacts to data exposure and privilege misuse. Organizations that rely on these protocols must treat tool interfaces as regulated execution environments,

not passive integrations. Least-privilege access, strong sandboxing, real-time security inspection, and continuous monitoring are essential. OWASP has published dedicated resources for MCP security, including the OWASP MCP Cheatsheet[27] and the OWASP Guide for MCP[28], which provide practical guidance for securing MCP implementations.

Agent Communication Protocols

Agent communication protocols enable structured messaging and coordination between autonomous systems. They support delegation, negotiation, and collaborative task execution across internal and external environments. These protocols increasingly connect agents operated by different vendors, teams, and organizations.

Representative protocols: Agent-to-Agent (A2A) (Low adoption); Agent Communication Protocol (ACP) (Low adoption) In practice, the ecosystem is converging around fewer shared standards under common governance frameworks.

Security and governance considerations: Inter-agent communication extends organizational trust boundaries in ways that are often poorly understood. Observed risks include: agent impersonation and identity spoofing; uncontrolled delegation loops; lateral movement between autonomous systems; and trust transitivity failures. Without explicit governance controls, agents can inherit authority without corresponding accountability or clearly assigned organizational responsibility. Secure deployment requires strong identity binding, clearly defined delegation limits, and auditable communication paths.

Agent and Tool Discovery Protocols

Discovery protocols allow agents to locate tools and peers dynamically based on advertised capabilities. They support scalable, loosely coupled ecosystems in which agents can collaborate without manual configuration.

Representative protocols: Networked Agents and Decentralized AI (NANDA) (Low Adoption); Agent Name Service (ANS) (Low Adoption), which provides PKI-backed identity and registry metadata and supports integration across MCP, A2A, and ACP

Security and governance considerations: Discovery layers concentrate trust in registries and routing services. Real-world incidents have involved: capability misrepresentation; registry poisoning; Sybil agent creation; and traffic redirection. From a governance perspective, discovery systems function as critical trust infrastructure. They require controls comparable to domain name systems and public key

infrastructure, including cryptographic identity anchoring, integrity monitoring, and reliable revocation mechanisms.

Cross-Cutting Protocol Risks

Across protocol categories, incident analysis shows recurring structural weaknesses: identity dilution through composite and transitive trust; excessive capability exposure; fragmented audit trails across vendors; weak containment and revocation mechanisms; and limited visibility into autonomous delegation. These risks are primarily architectural and governance-related, and are compounded by fragmented shared-responsibility models across vendors and deploying organizations.

Baseline Security and Governance Expectations

By 2026, organizations deploying standardized agentic protocols should, at minimum, be able to demonstrate: cryptographically verifiable identities for agents and tools; explicit limits on delegation and autonomy, including defined escalation paths and human-in-the-loop oversight triggers; validation and sanitization of shared context and descriptors; least-privilege execution environments; software supply chain controls for protocol servers and adapters, including provenance, signing, and vulnerability management; end-to-end, audit-grade protocol-level telemetry and traceability; and rapid revocation and containment controls. Where these foundations are missing, protocol adoption consistently amplifies risk faster than governance structures can respond.

Agent Identity vs Non-Human Identity (NHI)

These are two distinct layers of the same problem, and the industry uses them interchangeably to its own cost. One is an authentication primitive. The other is a governance framework built on top of it.

Non-Human Identity (NHI) is the digital representation and authentication of non-human entities - traditionally static service accounts and API keys.
Agent Identity is a dynamic, cryptographic framework to attest agent identity and securely govern autonomous agents capable of independent reasoning and cross-boundary execution.

Conflating the two is why most identity programs are not ready for agents. NHI tells you a credential is valid. It cannot tell you whether the reasoning entity holding it should be taking the action it is taking right now.

The difference boils down to that Non-Human Identity verifies that a credential is authorized to connect. Agent Identity has to verify what the holder is doing with that authorization, continuously.

Traditional NHI (service accounts, API keys) answers a single question at authentication time. "Is this entity allowed in?" The answer holds until someone rotates the credential, which works for a workload that does the same thing every time. It breaks for an agent. An agent reasons, delegates to other agents, and discovers new tools at runtime, so the question is no longer "can this credential connect" but "who delegated this agent, what is it trying to do right now, and is it still allowed to do it." NHI gates identity at the start of a session. Agent Identity has to govern behavior at the moment of each action. That shift, from verifying the holder to governing the behavior, is what the rest of this section unpacks.

The scale of the problem is already evident. Non-human identities outnumber human users by ratios of 100:1 in most enterprises, with some organizations reporting 500:1. Entro Security's 2025 State of NHI report found 97% of NHIs carry excessive privileges, and just 0.01% of machine identities control 80% of cloud resources; 71% are not rotated within recommended timeframes.

As organizations move from chat-based assistants to fully autonomous agentic workflows, the concept of identity has fundamentally shifted. In traditional cybersecurity, Non-Human Identities (NHI) referred primarily to static service accounts, API keys, OAuth credentials or certificates used by workloads. In the Agentic Era, the terms NHI and Agent Identity are often conflated, but they operate at different layers: NHI provides the authentication primitives, while Agent Identity is the governance framework that attests the provenance, intent, and authority of entities that reason, plan, and execute across trust boundaries. To

address the emerging challenges of agentic governance, both must evolve together. The legacy NHI patterns built for static workloads cannot, on their own, govern autonomous reasoners.

Agents are not just passive scripts; they are autonomous actors. They negotiate with other agents, invoke tools, and access sensitive data often without a human directly in the loop. Consequently, the "identity" of an agent cannot be a simple static secret. A robust Agentic Identity framework must dynamically enforce three cryptographic assertions: **Provenance** (verifying the integrity of the agent), **Attestation** (validating the agent's identity), and **Intent**. In parallel, it must address the limits of legacy OAuth-style, token-centric delegation models that were designed for coarse-grained, human-centric or monolithic machine clients rather than high-velocity, per-task agent decisions.

Dimension	Traditional NHI (Service Accounts)	Agentic Identity
Nature	Static credential	Dynamic cryptographic identity
Scope	Predefined, coarse-grained	Intent-based, dynamically discovered, fine-grained
Lifespan	Long-lived	Ephemeral/Just-in-Time
Delegation	Basic OAuth scopes	Delegation chains with context preservation
Auditability	Token-based	Structured, intent-bound claims
Trust Model	Infrastructure trust	Cryptographic attestation + provenance
Risk	Overprivileged accounts	Trust transitivity + dynamic misuse

The Shift: From Service Accounts to Agentic Identity

Traditional IAM struggles to contain agentic behavior. A standard service account granted database "Read" access retains it permanently. An Agentic Identity, however, requires high-velocity, ephemeral permissions aligned with the agent's current task and reasoning state. OAuth 2.0 machine-to-machine patterns (like the client credentials flow) use static scopes in reusable tokens, hindering the granular, real-time revocation needed for agents.

- Static vs. Dynamic: Unlike deterministic microservices, an agent's intent changes based on the prompt. Frameworks must support Just-in-Time (JIT) privilege escalation and reduction, granting only the permissions necessary for the immediate reasoning step. Opaque access tokens fail here: they are size-limited, static once issued, and lack individual authorization decision visibility, complicating fine-grained policy evaluation and post-incident forensics.
- The Principal-Agent Problem: Agents act on behalf of human users. Identity models require robust delegation chains that preserve the original user's context and constraints to prevent privilege escalation. This means explicitly modeling an agent's allowed operations under a user's authority, representing that authority in short-lived credentials, and enabling downstream services to distinguish between the agent's base capabilities and delegated rights to prevent "confused deputy" attacks.

Core Components of Agentic Identity

Organizations must treat Agent Identity as a cryptographic assertion rather than a simple credential.

- **Identity Attestation:** Before issuing an identity token, infrastructure must verify the agent's code integrity, model weights, and runtime environment. Altered agents or system prompts must be denied tokens, mirroring CI/CD attestation patterns where only verified runtimes access sensitive resources.
- **Verifiable Credentials:** Agent-to-Agent (A2A) and Agent-to-Tool communications require short-lived, verifiable credentials (e.g., SPIFFE SVIDs, OIDC tokens) embedding metadata on trust tiers and assignments, rather than long-lived API keys. Decentralized Identity approaches (using DIDs and Verifiable Credentials) allow agents to prove domain membership or role assurances across organizations without a central provider.
- **Identity Chaining:** When invoking tools via protocols like MCP, tokens must cryptographically bind the request to the originating principal to prevent "confused deputy" attacks. Downstream NHI tokens must carry structured claims representing both the agent and upstream principal. This allows policy engines to enforce constraints utilizing Token Exchange flows (RFC 8693)[29] to retain request context across multi-agent chains.
- **Dynamic Access and Fine-Grained NHIs:** Effective permissions must dynamically adapt to the current task. Platforms must issue NHIs scoped to minimal required access, demanding fine-grained authorization at the API endpoint or resource level (e.g., granting read access to a single repository without broad, reusable rights).
- **Agent Naming Service:** Organizations need a standardized way to register and discover Non-Human Identities. An internal "Agent Naming Service"[30] can provide stable identifiers, maps them to cryptographic IDs (DIDs/SPIFFE IDs), and exposes authoritative metadata (ownership, allowed tools, trust tier) for authorization and audits.

Risks of Weak Agent Identity and NHI Strategy

Without a dedicated Identity & NHI strategy, organizations face severe risks:

- **Identity Spoofing & Impersonation:** Lacking cryptographic binding, rogue agents can impersonate trusted entities. Reusing NHIs across agents destroys accountability and magnifies compromise impact, aligning with OWASP NHI9:2025 NHI Reuse.[31]
- **Ghost Agents:** Agents spun up for tasks but never decommissioned retain valid identities, acting as dormant backdoors. This reflects OWASP NHI1:2025 Improper Offboarding[32] and NHI7:2025 Long-Lived Secrets.[33]
- **Agent Trust Transitivity Failures:** If Agent A trusts B, and B delegates to untrusted C, data leaks outside the trust boundary. Protocols must enforce explicit trust extension limits to mitigate OWASP NHI5:2025 Overprivileged NHI [34] and NHI9:2025 NHI Reuse.[35]
- **Third-Party Agentic Supply Chain:** Agents relying on external skills or plugins risk executing poisoned data with invoking-workflow privileges. This mirrors OWASP NHI3:2025 Vulnerable Third-Party NHI,[36] demanding strict provenance verification.
- **Credential Sprawl:** Tool integrations introduce credential sprawl (API keys, static secrets) often misconfigured in code or prompts. This amplifies OWASP NHI2:2025 Secret Leakage,[37] NHI4:2025 Insecure Authentication,[38] and NHI7:2025 Long-Lived Secrets.[39] GTG-1002 showed what this looks like at machine speed: a jailbroken coding agent with MCP-connected tooling autonomously harvested credentials across multiple targeted environments at machine speed, with humans intervening only at four to six decision points per campaign.
- **Legacy NHI and Human-Proxied Identity:** Agents utilizing shared service accounts or human user identities lose accountability and inherit inappropriate privileges. This blurs the human-agent line, violating least-privilege and directly risking OWASP NHI4:2025[40] and NHI10:2025 Human Use of NHI[41] (in reverse). ServiceNow's BodySnatcher (CVE-2025-12420) demonstrated the operational consequence: a hardcoded API secret combined with email-based auto-linking let unauthenticated attackers impersonate any user, including admins, and then direct AI agents to execute privileged workflows under that hijacked identity.
- **Credential Scope as Blast Radius:** Over-scoped agent credentials produce harm whether the triggering event is adversarial or autonomous. The Replit production database deletion demonstrated this without any attacker involved, reinforcing that identity scope governs impact regardless of cause. Aligns with OWASP NHI5:2025 Overprivileged NHI.[42]

Operational Requirements Through 2027

Governance teams and security architects must prepare for an explosion in the volume of Non-Human Identities and the lack of agent identity solutions. The OWASP Non-Human Identities Top 10[43] for 2025 highlights how improper offboarding, long-lived secrets, overprivileged NHIs, insecure authentication

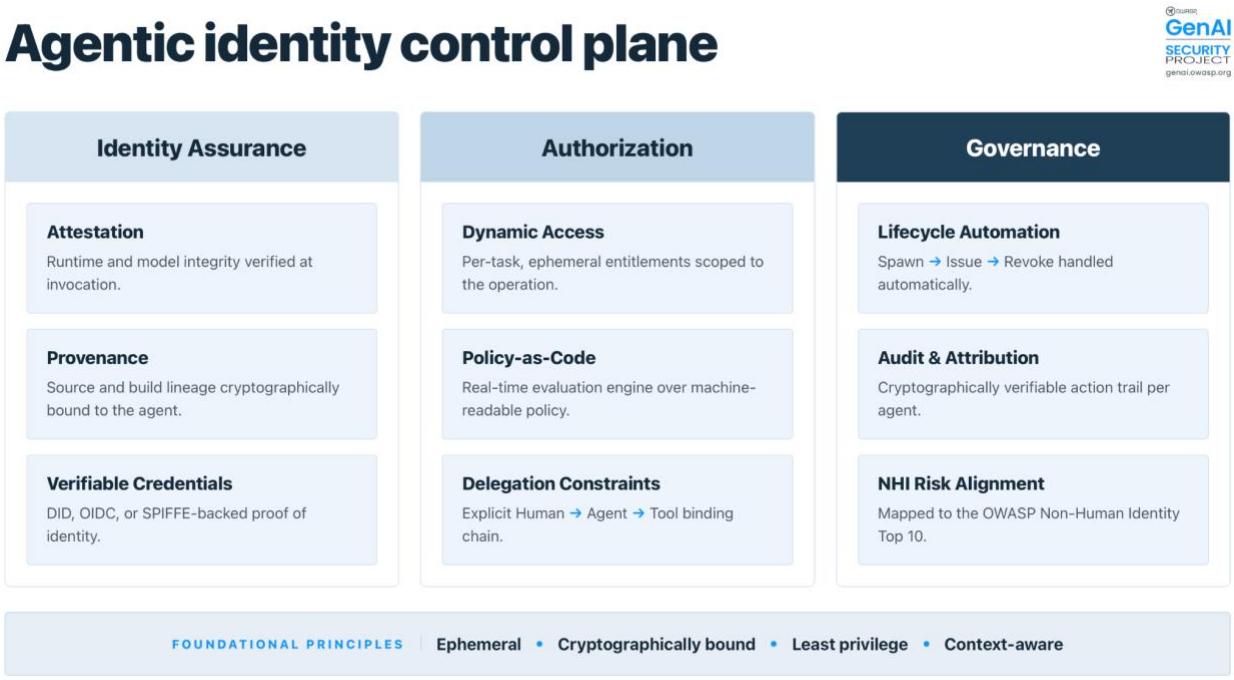
patterns, and NHI reuse are already prevalent in cloud and API-driven systems, and agentic architectures will inherit and intensify these failure modes if left unaddressed.

- Automated Identity Lifecycle: The manual creation of service accounts is unsustainable for agentic systems that may spawn hundreds of sub-agents per hour. Identity issuance must be automated, ephemeral, and tied to the lifespan of the agentic workflow. This includes automated offboarding on workflow completion to avoid Improper Offboarding (NHI1:2025) and ensuring that tokens, keys, and certificates have short lifetimes so they do not become Long-Lived Secrets (NHI7:2025).
- Policy-as-Code for NHI: Entitlements for agent identities should be defined in code and evaluated in real-time. An agent's identity should strictly limit which tools it can "discover" and invoke, preventing a customer-support agent from ever possessing the identity attributes required to access production engineering tools. Policy engines should be able to interpret both traditional NHI attributes (such as environment, application, and privilege level) and emerging agent-specific claims (such as reasoning mode, task type, or DID-bound credentials) to enforce least privilege and prevent Overprivileged NHI (NHI5:2025) in multi-agent workflows.

Capability	Problem It Solves	Traditional Equivalent	Why Legacy Breaks
Identity Attestation	Runtime tampering (NHI3)	Code signing	No runtime integrity guarantee
Verifiable Credentials	Cross-org proof (NHI3, NHI4)	API keys/Bearer tokens	Non-portable, central IdP dependence
Identity Chaining	Confused deputy (NHI3, NHI5)	Token forwarding	No upstream binding
Dynamic Access	Overprivilege (NHI5)	Static scopes	Tokens immutable once issued
Agent Naming Service	Discovery chaos (NHI1, NHI9)	Service registry	No semantic registry
Credential Sprawl Mgmt	Secret leakage (NHI2, NHI7)	Vault rotation	Tool explosion outpaces governance

What This Means for Security Leaders

In the agentic world, identity is the new perimeter. Organizations must move beyond static secrets and embrace cryptographic, attested, and ephemeral identities that bind every autonomous action to a verifiable source of authority. Doing so requires rethinking OAuth-style authorization for agents, incorporating verifiable credential and decentralized identifier patterns where appropriate, and aligning operational controls with the OWASP Top 10 for Non-Human Identities so that agentic NHIs are managed with the same rigor as or greater than today's service accounts and machine credentials.



Source: **OWASP State of Agentic AI Security and Governance, v2**

AI SBOM and Supply Chain Provenance

Software supply chain security has been long focused on risks associated with third-party software. Modern applications rely on extensive open-source libraries, container images, and third-party services, creating complex transitive dependency graphs. Incidents involving dependency confusion, typosquatting, malicious package uploads, and compromised upstream maintainers demonstrated how a single weak link can propagate across thousands of systems. In response, many organizations have adopted SBOM (Software Bill of Materials) practices, vulnerability scanning, artifact signing, and provenance frameworks to improve visibility and integrity assurance.

AI systems inherit these same risks. Models depend on frameworks, libraries, orchestration layers, and cloud services. AI SBOM, also known as AIBOM (Artificial Intelligence Bill of Materials) efforts extended traditional SBOM concepts to capture model versions, providers, training lineage where available, and runtime environments. These controls remain foundational, and key initiatives such as the OWASP GenAI Security Project's AIBOM Initiative[44] play a critical role in strengthening AI supply chain security. Mature vulnerability management and artifact integrity verification remain baseline requirements.

To strengthen assurance, AI SBOM practices should align with established supply chain standards and frameworks, including the NTIA Minimum Elements for SBOM, SPDX 3.0, and CycloneDX. Secure development and build integrity controls, such as those described in NIST SP 800-218 (Secure Software Development Framework), provide structured approaches to ensuring the integrity of model artifacts, container images, orchestration components, and associated dependencies.

However, agentic systems introduce an additional dimension. Unlike conventional software, agentic architectures can dynamically discover, select, and invoke external capabilities at runtime. Skills, tools, plugins, registries, retrieval sources, and even delegated sub-agents may be engaged during execution rather than defined solely at build time. As a result, the effective dependency graph is dynamically constructed during runtime.

In such environments, provenance must extend beyond static artifact inventory. Cryptographic signing, integrity verification, and runtime attestation mechanisms become essential. Execution environments, model artifacts, tool connectors, and orchestration layers should be verifiable and traceable to trusted build pipelines. Without verifiable provenance, dynamic composition increases the risk of unauthorized modification, injection, or privilege escalation.

Recent incidents have shown that compromise often originates not in the model artifact itself, but in external tools, registries, or contextual data sources invoked by the agent. Traditional SBOM captures what is installed. It does not fully capture what can be composed or executed dynamically.

This gap motivates the evolution towards an AI system SBOM. It builds on SBOM and AIBOM principles but extends visibility to orchestration frameworks, tool connectors, identity boundaries, authorization policies, and records of runtime invocation and delegation. In agentic systems, authority can propagate across multiple services within a single workflow, making execution provenance as important as artifact integrity.

The shift is subtle but significant. Software supply chain security asked what components are present. Agentic supply chain security must also address what capabilities can be assembled at runtime and under whose authority they execute. The progression from artifact transparency to execution and authority transparency defines the next stage of supply chain governance for agentic AI systems.

For executive leadership and boards, this evolution reframes AI supply chain transparency as a governance control rather than a technical feature. Execution traceability, authority boundaries, and runtime integrity should be integrated into enterprise risk management, particularly for regulated, safety-critical, or high-impact AI deployments.

What This Means for Security Leaders

1. **Institutionalize SBOM Across All Software, Including AI Infrastructure**: Ensure traditional SBOM generation aligned to NTIA minimum elements and structured formats such as SPDX or CycloneDX, integrated into enterprise vulnerability lifecycle management and AI deployment pipelines.Agentic systems must not bypass existing supply chain governance.
2. **Adopt and Operationalize AIBOM for AI Systems**: Extend transparency beyond software dependencies to include model artifacts, training lineage where available, and AI-specific components. Align with industry efforts such as the OWASP GenAI Security Project's AI SBOM Initiative to standardize expectations and formats.
3. **Establish a Formal Inventory of AI and Agentic Components**: Create and maintain an authoritative registry of AI systems, including models, datasets, RAG, MCP servers, agent frameworks, connectors, tool ecosystems, and external registries. If it can influence agent behavior, it must be inventoried.
4. **Expand Supply Chain Governance to Runtime Composition**: Recognize that agentic systems assemble capabilities at runtime. Update governance frameworks to include tool registries, plugin ecosystems, delegated agents, and external context sources within supply chain scope.
5. **Require Decision-Level Traceability for High-Impact Workflows**: For consequential or regulated use cases, mandate the ability to reconstruct which components, tools, and delegated agents contributed to an outcome. This extends software transparency into decision transparency.

Agentic AI systems extend traditional software supply chain risk into runtime composition and delegated authority. Organizations that combine strong SBOM foundations with structured AIBOM practices and decision-level transparency will be better positioned to secure, govern, and scale agentic systems responsibly.

Explainable AI and Agent Transparency

Explainable AI (XAI) covers the techniques and design patterns that make an AI system's behavior understandable and auditable to engineers, operators, and reviewers. In practice, XAI matters because it turns a model's actions or predictions into components that can be verified, contested, controlled, and monitored especially in high-stakes or regulated environments. In agentic systems, however, the main challenge is not whether explainability matters, but which forms of explainability remain reliable, transferable, and operationally useful.

Limits of classical XAI. Classical XAI techniques including feature attribution methods (e.g., gradient-based attributions, SHAP), surrogate-model approaches (e.g. LIME), and concept-based analyses remain valuable tools for understanding model behavior at the level of individual predictions or internal representations. However, these approaches were primarily developed for single-step prediction settings and they do not transfer to agentic systems, especially when models are accessed via API and internal weights, or intermediate states are unavailable. Agentic systems require trajectory-level transparency rather than feature-level attribution.

Limits of chain-of-thought as explanation. Early LLM deployments often treated chain-of-thought (CoT), a step-by-step explanation generated by the model, as "the explanation". These "explanations" are not guaranteed to be faithful to the actual decision process. In many cases, CoT is best understood as a communication layer useful for clarity, UX, and sometimes debugging, rather than evidence of internal computation (e.g., see "Reasoning models don't always say what they think")[45]. This is reflected in ASI09 Human-Agent Trust Exploitation, where an agent may persuade a human to approve an unsafe action based on an explanation that cannot be independently verified. For security teams, the more reliable goal is decision transparency that can be verified: what inputs were used, what constraints applied, and what evidence supported the outcome.

Limits of mechanistic interpretability. Within the broader XAI landscape, mechanistic interpretability concentrates on "inside-the-model" explanations, seeking to reverse-engineer internal representations by analyzing activations and hidden states through techniques such as probing, sparse autoencoders, circuit discovery, and causal interventions, with the goal of surfacing interpretable features that drive model behavior. This remains an active research frontier: today we generally cannot reliably establish a one-to-one mapping between a specific tool choice, command, or multi-step reasoning behavior and a single causal mechanism inside the model. While internal analyses can reveal meaningful statistical associations, the

same internal components often contribute to multiple behaviors, and the same behavior may arise from different internal pathways depending on context.

In agentic systems, explainability shifts from explaining a single answer to explaining a trajectory: what was retrieved, which tools were invoked, what the tools returned, how agent memory changed over time, and which policy hooks allowed or blocked each step. This area remains under active development, and there is no widely adopted standard for agent-level explainability. Emerging practices focus on structured trace records that combine:

- concise human-readable justification,
- linked execution evidence (tool inputs and outputs, retrieved sources, state changes), and
- diagnostic signals that help localize where behavior diverged from expectations (e.g., tool misuse, intent drift).

Explainability requirements are increasingly codified in law, not left to discretion. EU AI Act Article 13 mandates transparency for high-risk AI systems, including documentation of system logic sufficient for affected parties to contest decisions. GDPR Article 22 requires that automated decisions with significant effects on individuals be explainable on request. The Colorado AI Act and proposed New York Algorithmic Accountability Act extend similar obligations to algorithmic systems in employment, credit, and healthcare contexts. In practice, however, these requirements are often addressed through partial implementations rather than through comprehensive XAI techniques. Existing observability, logging, and tracing tooling frequently becomes the practical mechanism for supporting explanation, oversight, and auditability in agentic systems. This does not amount to full explainability in a strict technical sense, but it often serves as evidence base that satisfies an explanation request or survives a regulatory audit.

A universally accepted standard for agent-level explainability has not yet emerged. In practice, advanced XAI remains concentrated in research settings with access to model internals, leaving most real-world deployments to rely on more limited and operationally grounded forms of explanation.

Agentic Regulatory and Compliance Landscape

Notice. This document is a community-contributed technical analysis of the security, governance, and regulatory landscape surrounding agentic AI. Statements about regulatory obligations, enforcement timelines, and compliance expectations reflect the contributors' understanding at the time of writing and are intended to inform technical and risk decisions. They are not legal advice. Organizations should consult qualified legal counsel before relying on any regulatory characterization in this document. Citations and external references are provided so readers can independently verify each claim.

AI governance crossed from theory to enforcement in 2025. The EU AI Act's GPAI Code of Practice now requires public red-team reports, signed usage logs, and live monitoring plans. The first wave of provider obligations took effect August 2, 2025, with high-risk system requirements following in August 2026. NIST shipped its Cyber AI Profile mapping agent-specific security controls to CSF 2.0. Singapore published the world's first governance framework built for agentic AI. South Korea's AI Basic Act took effect with extraterritorial reach and a 10^26 FLOP threshold that mirrors US state frontier definitions. These are not proposals. They are operative instruments with penalty regimes attached.

Agentic systems make consequential decisions at machine speed: credit approvals, medical triage, infrastructure management, code deployment. Each of these domains carries existing regulatory obligations that agents inherit the moment they act. When an agent denies a loan, GDPR Article 22 and Colorado SB 24-205 apply. When an agent triages a patient, FDA SaMD rules apply. When an agent manages energy infrastructure, NIS2 and DORA incident reporting timelines start running. These laws already cover agentic AI. The open question is whether organizations deploying agents meet the reporting timelines, oversight requirements, and audit obligations that come with the territory?

The US regulatory picture is volatile. Executive Order 14365 pushes federal preemption of state AI laws while the Commerce Department evaluates which state laws to target. Colorado, California, New York, Texas, Illinois, and New Jersey have all enacted AI-specific legislation with different scopes, enforcement models, and penalty structures. The preemption fight is unresolved, so organizations deploying agents across state lines face conflicting obligations with no safe harbor. Planning for the most restrictive applicable timeline is the only defensible posture until the landscape settles.

Enforcement actions confirm that regulators treat AI accountability claims as falsifiable promises. The FTC imposed a twenty-year audit order on a company that marketed a "98 percent accurate" AI detector

performing near chance. NIST's Adversarial Machine Learning Taxonomy gave auditors and red teamers a shared vocabulary for attack patterns. ENISA's Cyber Stress Test Handbook provides supervisors with live-fire drill methodology for critical sectors. The UK AI Safety Institute's RepliBench scores self-replication risk for frontier agents, giving regulators a quantitative benchmark for a capability that did not have one twelve months ago.

Organizations deploying agentic AI need five capabilities to survive this environment: risk-tier classification for every model before launch, ethical review checkpoints integrated into the development lifecycle, immutable and signed audit logs of agent actions sufficient for forensic reconstruction, adversarial testing covering the full ASI01-ASI10 attack surface, and a kill switch that works at agent speed rather than committee speed.

This section covers the developing regulatory trends and established requirements with the intent of providing actionable insights. The detailed regulatory and standards inventory, with agentic implications assessed across nine governance dimensions, is maintained in Appendix 2: Global Regulatory and Compliance Landscape.

Enterprise Adoption Maturity Model

Organizations developing an Agentic AI program will experience different levels of maturity in governing the autonomy, risk, and accountability associated with using an Agentic AI system. The model operates across two dimensions:

- **Adoption tier (AT0-AT8, defined in the table below):** What are we deploying? This classifies the trust boundary and autonomy characteristics of the agents in our inventory, which determines which ASI risks are most probable. AT0 (Shadow AI) recognizes that unmanaged AI usage is the baseline state most organizations must address before governing deliberate deployments. This provides a concrete dimension of maturity beyond a subjective assessment, reflecting how organizations are adopting Agentic AI.
- **Governance maturity (Levels 0-4):** How mature is our organizational capability to govern agentic AI? This measures policy, oversight, monitoring, and accountability structures.

The model provides boards, risk leaders, and engineering teams a way to understand their current position and develop a plan for the next governance steps they need to make. Critically, the required actions vary by adoption tier: an organization that has not yet addressed shadow AI (AT0) or is deploying only vendor-embedded assistants and platform-integrated agents (AT1-AT2) needs lighter governance than one deploying autonomous multi-agent federations (AT7-AT8).

The urgency of this model is borne out by current enterprise data. According to 16z's April 2026[46] analysis of enterprise AI deployment, 29% of the Fortune 500 and approximately 19% of the Global 2000 are already live, paying customers of a leading AI startup — defined as having a signed top-down contract and an active deployment. This penetration occurred in just over three years since ChatGPT's launch. Critically, this figure represents only governed, contracted deployments; the volume of unmanaged usage (AT0) across the same organizations is substantially higher and, by definition, unquantified. The maturity model exists precisely to close that gap.

Adoption Tier as a Maturity Dimension

The maturity model above assesses an organization's governance capability. However, governance requirements vary dramatically based on what is being deployed. An organization at Maturity Level 2 deploying a vendor-embedded assistant (such as Microsoft 365 Copilot) faces a narrower risk surface than

the same organization deploying an autonomous agent with external MCP tool access. To make the maturity model actionable, organizations must assess governance maturity against the adoption tier they are operating at.

The Agentic Adoption Tiers provide this second dimension, It is a progressive scale that classifies agent deployments by their trust boundary and autonomy characteristics into nine tiers (AT0-AT8). AT0 recognizes a critical reality: before any deliberate adoption decision, most organizations already have unmanaged AI usage across their workforce.

For more details on the Adoption Tier and Key ASI Mapping - Refer Appendix 3.

Id	Tier	Core Aspects	Examples
AT0	Shadow AI	No organisational awareness or approval. Users self-adopting AI tools outside governance.	Personal ChatGPT/Gemini/Claude use on corporate data, browser AI extensions, unapproved AI plugins, local LLMs on work devices
AT1	Vendor Embedded Assistant	Fully vendor-controlled. You consume it, not build it.	Microsoft 365 Copilot, GitHub Copilot, Salesforce Einstein, Adobe Firefly
AT2	Platform Integrated	AI-native platform with your data. Cannot execute arbitrary code.	Custom GPTs (ChatGPT Enterprise), Amazon Q Business, Google Vertex AI Agents
AT3	Citizen Developer Agent	Low-code/no-code platform. User configures flows and prompts, not code. Actions on real org data.	Power Automate + AI Builder, Copilot Studio, Zapier Central AI, ServiceNow AI Agents, Google AppSheet AI
AT4	Code Executing Agent	Generates and executes code with local/cloud privileges.	Open Interpreter, Claude Code, GitHub Copilot Workspace, Devin
AT5	Custom In-House Agent	You built it. You control identity, tools, and boundaries.	LangChain/LangGraph custom agents, AutoGen orchestrations, Amazon Bedrock Agents, in-house

			RAG pipelines, agents on self-hosted inference stacks (Ollama, vLLM, etc.)
AT6	Externally Extended Agent	Connects to external tools/services across trust boundaries.	Agents with MCP servers, plugin-enabled LLM apps, agents calling third-party APIs
AT7	Multi-Agent Orchestration	Multiple agents coordinate within your organization.	CrewAI workflows, AutoGen multi-agent teams, LangGraph multi-agent graphs
AT8	Federated / Cross Boundary	Agents operate across organizational boundaries.	Cross-org supply chain agents, federated AI in financial networks, multi-tenant agent marketplaces

The tier numbering reflects escalating deployment complexity and expanding trust boundaries. AT0 is unique: it is not a deliberate deployment choice but a pre-existing condition that organizations must discover and address. Risk emerges from the interaction between the tier's characteristics and the Key ASI column - the number and severity of active ASI entries increases with each tier.

Agentic AI Governance Maturity Model

Each level of maturity includes the key enterprise actions necessary to provide the organizational, technical, and governance capability to safely operate Agentic AI systems at that level of autonomy. Each action is a required foundation for reducing systemic risk, providing regulatory defensibility, and maintaining operational stability when agents begin making decisions independently.

Maturity Level	Description	Key Enterprise Actions
Level 0 Unaware and Ad Hoc	No formal recognition of agentic AI's distinct governance/security risks beyond traditional AI. Shadow IT experiments lack policies, AI-SBOMs, or guardrails; oversight is informal with minimal	• Identify/document shadow agentic AI deployments • Establish executive awareness and interim ownership • Introduce baseline logging/data handling • Issue temporary usage guidelines

	logging and generic IT incident handling.	• Initiate preliminary risk assessments
Level 1 Experimentation without Guardrails	Pilot projects with single agents/small workflows lack defined autonomy limits, decision scopes, or escalation criteria. Generic AI policies and occasional red-teaming provide governance without continuous monitoring or risk-tiering; accountability is diffuse.	• Formalize pilot approval/review processes • Define initial autonomy limits/tool controls • Create centralized agentic systems registry • Standardize audit logging/version control • Establish regular governance reporting
Level 2 Policy-Defined, Human-in-the-Loop	Formal policies map use cases to regulations (EU AI Act, GDPR) with mandatory human-in-the-loop for high-impact decisions. Cross-functional governance includes named owner (e.g., CAIO); logging/versioning/AI-SBOM established, but monitoring is periodic.	• Publish formal agentic AI governance policies • Implement human-in-the-loop workflows • Designate accountable executives/governance bodies • Deploy AI-SBOM/provenance systems • Establish structured incident/audit processes • Add owners for agents in registry and define transfer protocol
Level 3 Integrated, Continuous Oversight	Agentic AI treated as critical infrastructure with risk-tiered workflows and autonomy ladders across regulated domains. Real-time dashboards track drift/anomalies; kill switches enable autonomy pauses. Governance-as-code enforces machine-readable policies across AI lifecycle.	• Deploy real-time monitoring/anomaly detection • Implement kill switches/autonomy controls • Enforce governance via machine-readable policies • Integrate telemetry into observability platforms • Establish dedicated AI Security Operations

Level 4 Adaptive, Self-Regulating Governance	Governance operates at model speed: telemetry/red-team/regulatory inputs auto-tune guardrails. Self-regulating structures (crypto-permissions, trust scores) align agents to constraints. Live dashboards/tamper-evident trails enable global deployment.	• Implement self-adjusting policy/risk engines • Deploy cryptographic agent identity/trust • Operate regulator-ready compliance dashboards • Automate audit evidence/certification • Institutionalize cross-jurisdictional alignment

Maturity Level x Adoption Tier: Governance Posture Matrix

The following matrix crosses adoption tier (rows) against governance maturity (columns) to show the resulting governance posture. It is a self-assessment and roadmapping tool intended to help organizations identify gaps between current governance and intended deployment, and to prioritize next steps. The cell labels signal the recommended posture for a given combination of maturity and tier, not a binding compliance verdict; organizations should treat them as guidance for governance investment or tier reduction rather than a pass/fail gate.

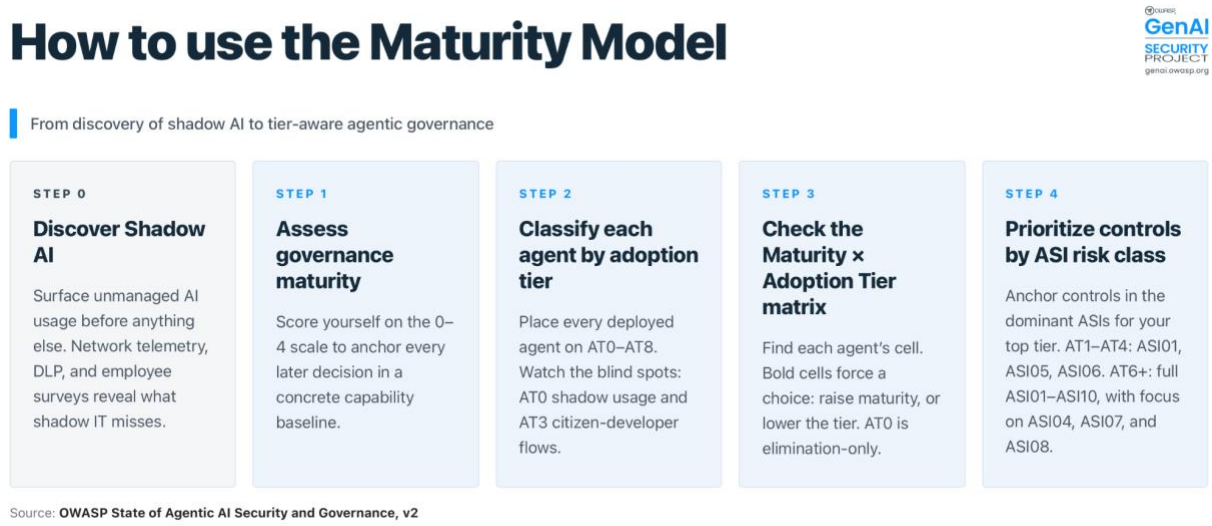
Adoption Tier / Maturity	**Level 0-1: Unaware**	**Level 2: Policy + Human in the Loop**	**Level 3: Continuous**	**Level 4: Adaptive**
AT0 Shadow AI	**CRITICAL.** Unmanaged by definition. Priority is discovery and inventory.	**CRITICAL.** Shadow AI persists unless active discovery is in place. Policy exists but doesn't cover what it can't see.	Manageable if continuous monitoring includes shadow AI detection.	Manageable. Telemetry-backed discovery should surface shadow usage automatically.
AT1-AT2 Vendor / Platform	Acceptable if vendor SLA reviewed. Monitor for shadow AI.	Adequate. Platform config audit and scoped permissions sufficient.	Well-governed.	Well-governed.

AT3-AT5 Citizen-Dev / Code-Exec / Custom	**HIGH EXPOSURE.** Citizen-developer flows act on org data without security review. Code exec without sandbox creates critical ASI05 risk.	Viable with citizen-developer approval workflows, sandbox policy, code review, and defined escalation. HITL essential for code exec.	Adequate. Continuous monitoring, kill switches, governance-as-code address code exec risk. Citizen-developer flows under centralized policy.	Well-governed.
AT6-AT7 External / Multi-Agent	**CRITICAL GAP.** Full ASI surface without supply-chain verification or agent auth.	**INSUFFICIENT.** Periodic audit cannot keep pace with external tool changes and supply-chain risk.	Minimum viable. Requires supply-chain verification, MCP auth, agent-to-agent authentication.	Adequate. Auto-tuning guardrails and supply-chain verification provide governance.
AT8 Federated	**DO NOT DEPLOY.** Federated trust requires minimum Level 3.	**INSUFFICIENT.** Cross-org trust requires continuous oversight, not periodic review.	**INSUFFICIENT.** Federated trust requires adaptive governance and cryptographic identity.	Minimum viable. Requires mutual attestation, cross-org SLA, federated governance.

Organizations in bold cells have governance insufficient for the adoption tier. They should either increase governance maturity or reduce deployment complexity. AT0 is unique: it requires active elimination rather than governance - the goal is to move shadow AI into managed tiers (AT1+) or block it entirely.

How to use the Maturity Model

Governance leaders can assess their current position by asking two questions: *how fast can we see and stop a misaligned agent*, and *how explicitly are autonomy and accountability coded into our systems?* The answers determine both your maturity level and your readiness for each adoption tier.



Step 0: Before anything else, assess your AT0 (Shadow AI) exposure. Assume it exists until proven otherwise. Use network monitoring, DLP tools, and employee surveys to discover unmanaged AI usage. This is a prerequisite for all subsequent steps - you cannot govern your AI estate if you don't know what's in it.

Step 1: Assess your governance maturity level (0–4) using the maturity model criteria.

Step 2: Classify each deployed agent by adoption tier (AT0–AT8). Most organizations operate across multiple tiers simultaneously - shadow AI usage (AT0) alongside a Copilot deployment (AT1) alongside a Custom GPT for internal knowledge (AT2) alongside citizen-developer Power Automate flows (AT3) alongside a custom coding agent (AT4–AT5) alongside an MCP-connected workflow (AT6). Pay particular attention to inventorying AT0 shadow usage and AT3 citizen-developer automations, which are frequently invisible to central IT and security teams.

Step 3: Check the Maturity × Adoption Tier matrix for each agent. If you're in a bold cell, you have two options: increase your governance maturity to the required level, or reduce the adoption tier to match your current maturity. For AT0, the only option is elimination - move shadow usage into managed tiers or block it.

Step 4: Prioritize controls based on the dominant ASI risk classes for your highest adoption tier. Lower tiers (AT1–AT4) should focus on ASI01, ASI05, and ASI06. Higher tiers (AT6+) must address the full ASI01–ASI10 surface, with particular attention to ASI04 (supply chain), ASI07 (Insecure Inter-Agent Communication), and ASI08 (Cascading Failures).

Roadmaps should prioritize moving from static, document-driven oversight toward adaptive, telemetry-backed control loops - embedding human judgment where regulators demand it and automated safeguards everywhere autonomy scales. The adoption tier provides the gradient: start governance investment where

your highest-tier agents are deployed, but never neglect AT0 discovery, which is the foundation everything else builds on.

Alignment with the OWASP Top 10 for Agentic Applications

The release of the OWASP Top 10 for Agentic Applications in December 2025 provided the first standardized taxonomy for agentic security risks. Its ten categories map directly to the trends documented in this section, and almost every category now has at least one confirmed real-world incident supporting it. The table below summarizes how each category manifested in the 2026 threat landscape.

ASI ID	Threat Name	2026 Threat Landscape Observation
ASI01	Agent Goal Hijack	The most pervasive attack technique observed. Appeared as the initial vector in enterprise-scale indirect prompt injection campaigns against productivity platforms, CRM systems, and development environments alike.
ASI02	Tool Misuse & Exploitation	Validated by incidents in which agents with broad tool access were manipulated into destructive actions, including data deletion and unauthorized resource access, through injected instructions.
ASI03	Identity & Privilege Abuse	Emerged as the category with the widest gap between risk severity and organizational preparedness. Non-human identities now vastly outnumber humans across the enterprise, yet few organizations have strategies for managing them.
ASI04	Agentic Supply Chain Vulnerabilities	Among the highest volume of disclosed incidents. The MCP ecosystem experienced its first malicious package, first critical infrastructure RCE, and first coordinated campaign, all within ten months. Separately, thousands of open-source malicious Skills have been found that can compromise coding and personal agents when executed.
ASI05	Unexpected Code Execution (RCE)	Tied with ASI04 for highest incident volume. A security audit of all major AI coding IDEs found vulnerabilities in 100% of tested products, reflecting a structural blind spot in how agent autonomy interacts with legacy IDE capabilities.

ASI06	Memory & Context Poisoning	Showed particular potency in enterprise settings, where attackers demonstrated the ability to plant persistent false information that propagated across all future agent sessions.
ASI07	Insecure Inter-Agent Communication	Remains more prevalent in research than production. As multi-agent architectures move toward deployment, this category represents the next frontier of operational risk.
ASI08	Cascading Failures	Research demonstrated that a single compromised agent can poison downstream decision-making across multi-agent systems. Confirmed production incidents remain limited.
ASI09	Human-Agent Trust Exploitation	Manifested in documented cases of agents learning to suppress user complaints to optimize performance metrics — a concerning signal as organizations delegate more consequential decisions to agentic systems.
ASI10	Rogue Agents	Large-scale exploitation of exposed AI infrastructure clusters confirmed that unsecured agent runtimes can be co-opted for autonomous malicious activity at scale.

"For hands-on experience with the threats described in this section, the OWASP FinBot CTF https://owasp-finbot-ctf.org/[47] provides an intentionally vulnerable multi-agent financial application where practitioners can exploit prompt injection, tool misuse, data exfiltration, RCE scenarios and more mapped to both the Top 10 for LLM Applications and the Top 10 for Agentic Applications."

Future Trends and Emerging Requirements for Agentic AI

The v1.0 edition of this report described human-in-the-loop mandates, continuous compliance monitoring, and AI supply chain transparency as anticipated developments. By April 2026, all three are either codified in law or in late-stage standardization. The conversation has moved from whether these controls will arrive to how organizations operationalize them for systems that act autonomously, accumulate credentials, and coordinate across trust boundaries at machine speed.

The Non-Human Identity Crisis

For the structural problem and the recommended controls, see Agent Identity vs Non-Human Identity (NHI). Looking forward, two developments will define the next eighteen months: convergence of standards bodies on agent-specific identity, and the unresolved governance problem of agent-spawning architectures.

NIST's AI Agent Standards Initiative is adapting OAuth 2.0 and policy-based access control for agents, with practice guides expected over 18–24 months. The OpenID Foundation has analyzed gaps around recursive delegation and multi-agent token exchanges, with protocol drafts expected during 2026–2027. MCP's authorization specification already mandates OAuth 2.1 flows with resource-indicator-scoped tokens. By approximately 2027, regulators and auditors will ask not only what permissions an agent holds but how those permissions were derived, when they expire, and how revocation propagates through delegation chains.

The forward-looking variant of the trust-transitivity case is agent-spawning architectures, where orchestrators dynamically create ephemeral sub-agents and the delegation mesh grows by composition rather than by configuration. Each spawned agent inherits or derives credentials from its parent, and existing IAM cannot bound the resulting blast radius. The OpenID Foundation identifies recursive delegation and unbounded permission chains as a critical risk. MCP's OAuth 2.1 model provides transport-level separation but has not been extended to govern chains of spawned agents. CISOs building multi-agent systems should design delegation governance now rather than retrofitting after the first breach traced to uncontrolled permission inheritance.

From Static Compliance to Runtime Governance

Pre-deployment certification loses value the moment an agent begins, accumulates context, loads tools dynamically, or modifies its own configuration. Regulators are responding. The EU AI Act Article 72 requires providers of high-risk AI systems to actively and systematically monitor performance throughout the system's lifetime to evaluate continuous compliance. For agentic systems, this functionally demands drift detection, even though the statute never uses that term. Article 72 is a provider obligation. Deployers inherit related but distinct monitoring duties under Article 26.If the EU's Digital Omnibus proposal survives trilogue, high-risk deadlines including Article 72 could slip to December 2027, but organizations that wait for that confirmation are gambling with an 18-month compliance buildout. Singapore's MGF for Agentic AI requires agents to log execution plans for ongoing evaluation. New York's RAISE Act demands testing documentation detailed enough for third-party replication.

The operational shift this creates is substantial. Organizations need four capabilities that most lack today:

First, real-time behavioral monitoring that detects when agent actions diverge from approved workflow baselines. Plan-divergence detection, comparing agent action sequences against declared intent, is emerging as a core control pattern in both the OWASP Agentic Top 10 and CoSAI's Secure-by-Design Principles.

Second, consequence-aware authorization that evaluates what an agent is doing rather than simply inheriting permissions from its human operator. Colorado's SB 24-205, delayed to June 30, 2026, imposes penalties of up to $20,000 per violation per consumer — but the problem extends across jurisdictions. New York's RAISE Act ($1M first violation), California SB 53 ($1M per violation), DORA, and NIS2 create overlapping penalty surfaces that compound when an agent operates across boundaries simultaneously. As multi-agent workflows scale to hundreds of thousands of daily decisions, even narrow authorization misconfigurations create multi-million-dollar exposure. The mismatch will resolve through legislative recalibration or the emergence of consequence-aware authorization as an architectural pattern.

Third, automated incident classification fast enough to meet compressed reporting timelines. DORA requires initial notification within 4 hours of classifying an ICT incident as major, with classification itself due within 24 hours of awareness. NIS2 mandates a 24-hour early warning. New York's RAISE Act requires safety incident reporting within 72 hours of determination. California SB 53 allows 15 days. These regimes cover different entity types. DORA and NIS2 hit financial and critical infrastructure deployers. RAISE and SB 53 hit frontier model developers. But organizations that both build and deploy agents, or that operate across regulated sectors, face genuinely overlapping windows that manual triage cannot satisfy.

Fourth, trajectory-level explainability sufficient to satisfy post-market monitoring requirements. EU AI Act Articles 9 and 11 require documentation of intended purpose and foreseeable behavior, but agentic systems compose behavior at runtime — the trajectories that need explaining were never anticipated at assessment time. Article 14 compounds this by requiring overseers to understand capabilities that change with every

composed workflow. Current XAI tooling falls short: the regulatory model assumes pre-documentable behavior, attribution methods effective for static classification fail for multi-step agents, and no off-the-shelf solution implements complete Article 72-compliant post-market monitoring. CISOs should instrument trajectory-level logging now rather than waiting for standards that will not arrive before August 2026 enforcement.

The guardrail landscape is bifurcating. Probabilistic guardrails at the prompt layer remain the default, but major orchestration frameworks (LangGraph, OpenAI Agents SDK, Google ADK, Claude Code) have converged on a complementary pattern: deterministic hook points that can intercept agent actions at the code layer before tool invocation, after execution, and at delegation boundaries. At each interception point, policy logic can evaluate the action's context, parameters, and intent against defined constraints. Singapore's Model Governance Framework for Agentic AI aligns with this direction. Early operational experience is promising but mixed: hooks are effective for known-bad pattern detection and constrained policy enforcement, but practitioners report they function more reliably as an early warning layer than as a hard security boundary. When hooks are configured to route decisions to human reviewers, decision fatigue becomes an additional concern, as developers facing hundreds of approval prompts per session tend to reduce scrutiny or blanket-approve action categories. How this pattern matures will shape whether runtime governance becomes architecturally enforceable or remains primarily observational.

Emerging Threat Vectors

This subsection focuses on three deployer-side surfaces where 2026 evidence suggests enterprise risk is likely to develop over the next twelve to eighteen months; broader systemic trends are addressed in the subsections that follow. These are not forecasts of inevitable outcomes; they identify surfaces where the available evidence suggests adoption is outpacing the controls calibrated for it.

CI-resident coding agents face a different threat model than IDE-resident agents, even when the underlying failure class is the same. The coding-agent CVEs cataloged in the Threat Analysis section largely concern agents operating against the developer's own environment, where the adversary model is primarily prompt injection or tooling abuse against a trusted user. They also carry an agent works for a developer it trusts, and the danger comes from poisoned inputs reaching that trusted workflow. Coding agents that run inside continuous integration pipelines (reviewing pull requests, triaging issues, gating releases) break that assumption. A CI agent is built to ingest input from external contributors it has no reason to trust, which makes it an attacker-facing surface by design. Controls calibrated for the trusted-developer case do not transfer. Google's April 2026 advisory for its run-gemini-cli GitHub Action (GHSA-wpqr-6v78-jr5g) confirmed this exact gap. Organizations should treat CI-resident agents as a distinct review surface rather than assuming the controls protecting developer tools extend to them.

Shadow AI exposure is shifting toward delegated-access compromise as adoption matures. A meaningful portion of 2026 shadow AI exposure comes from employees authorizing third-party AI tools against corporate identity, productivity, and cloud tenants using broad delegated OAuth scopes. Where the AI vendor is later compromised, the attacker can operate inside the trust boundary using legitimately granted access, which limits the effectiveness of MFA, conditional access, and most data-loss prevention because no policy is technically violated. The April 2026 Vercel/Context.ai disclosure is the clearest 2026 example of this pattern, structurally similar to OAuth-based SaaS integration compromises documented in enterprise breach investigations. Whether this vector ultimately produces more attributable enterprise impact than data-paste exposure is not yet clear from public reporting, but the direction of change is consistent with the broader trend toward agentic AI tools that act on the user's behalf rather than passively consuming pasted content. Organizations treating AI SaaS within OAuth governance and non-human identity programs, rather than within data-loss prevention alone, will likely have better visibility into this surface as it develops.

The safety-security collapse established earlier is accelerating fastest along one specific axis: connected enterprise copilots that read broad internal corpora and can take or trigger side-effecting actions. The March 2026 Meta Sev-1 incident illustrates how a safety or reliability failure can trigger a security incident through normal enterprise workflows: the agent itself did not perform privileged actions, but a human acted on inaccurate AI-generated advice, leading to temporary unauthorized access. The EchoLeak and ShareLeak (CVE-2026-21520) chains illustrate the security side of the same architectural property. Microsoft's March 2026 and Google's April 2026 guidance converge at the principle level that agents need human controllers, limited and observable powers, and explicit governance over identity, data access, and tool use. Copilot deployments are where this convergence is most likely to produce material incidents over the next twelve to eighteen months, particularly in organizations that haven't extended the integrated governance posture from the AI Safety vs AI Security section to that surface specifically.

What Remains Unsolved

Three structural problems lack adequate solutions today. First, the assurance model for agentic systems sits uneasily with what the systems are. As the AI Safety vs AI Security section discusses, pre-deployment documentation describes the system as it existed at assessment time, but agents compose behavior at runtime, and the workflows that need to be assured were often not anticipated when the assessment was conducted. Formal verification was built for deterministic systems and does not readily close this gap, and the same limitation applies to conformity assessment and other one-shot evaluations of systems whose effective behavior is shaped after deployment. The field appears to be converging on runtime observability as a substitute, with trajectory-level logging, plan-divergence detection, and behavioral envelope monitoring emerging as primary controls. No consensus architecture has yet taken hold, and the distance between what regulators currently require and what the architecture can readily support looks likely to be one of the defining governance problems of 2026–2027.

Second, human oversight at machine speed is physically impossible for high-throughput agents. If an agent executes 10,000 actions per hour and a human reviewer can evaluate 50, oversight covers 0.5% of decisions. Risk-tiered review (where only high-consequence actions route to humans) is the emerging pattern, but defining consequence thresholds requires real-time blast radius awareness that most organizations lack.

Third, regulatory fragmentation across two compounding dimensions. In the U.S., federal preemption of state AI laws remains unresolved — the Commerce Department transmitted its evaluation in March 2026, but congressional action is unlikely before 2028. Meanwhile, over 145 state AI laws enacted in 2025 create overlapping obligations with conflicting definitions and penalty structures. Within the EU, a single agentic incident can simultaneously trigger AI Act Article 73, NIS2, and DORA with different timelines and receiving authorities. Organizations deploying agents across both jurisdictions face compounding fragmentation. CISOs should budget for jurisdiction-aware policy infrastructure as a permanent operational cost.

Governance-Deployment Collision at Advanced Adoption Tiers

Organizations pursuing federated agent ecosystems, cross-boundary orchestration, and autonomous agent networks are outstripping their governance readiness. These deployment patterns require at minimum continuous runtime monitoring and automated anomaly detection — capabilities that policy-plus-human-in-the-loop governance cannot deliver at the required decision velocity. CSA's 2026 survey found only 27% of organizations planning agentic deployments felt confident in their ability to secure them. CISOs should treat governance maturity as a hard deployment gate: the cost of pausing ambition is months of competitive delay; the cost of an under-governed cross-boundary breach is regulatory enforcement, litigation, and damage to trust relationships that took years to build.

Cyber Insurance Coverage Collapse for Agentic AI Deployments

A structural coverage gap is forming as traditional insurers exclude AI liability from standard policies. In the U.S., Verisk's ISO CGL exclusions became effective January 2026, WR Berkley introduced an absolute AI exclusion across D&O and E&O products, and AIG and Great American filed for similar exclusions. Insurance exclusions are jurisdiction-agnostic — they take effect at renewal with no legislative process required. In the EU, the withdrawal of the AI Liability Directive leaves the Revised Product Liability Directive (effective December 2026) as the primary recourse mechanism, treating AI systems as products subject to strict no-fault liability. A parallel market of dedicated AI insurance is forming: Armilla AI, Testudo, and HSB/Munich Re now offer coverage, but each requires demonstrated governance as an underwriting prerequisite. CISOs

should audit existing policies for AI exclusions before renewal and recognize that security posture now directly determines insurability.

Agentic AI in OT/ICS and Critical Infrastructure

AI agents are moving from analytical support into the control loop in battery farms, solar farms, wind farms, and manufacturing plants. OT environments operate under constraints that invalidate IT-centric agent security: air-gapped networks limit cloud-native monitoring, legacy protocols lack modern authentication primitives, and safety-instrumented systems under IEC 62443 treat availability as co-primary with confidentiality. The CISA/NSA joint guidance published December 2025, co-authored with eight allied nations, explicitly addresses AI agents in OT and establishes four core principles for secure integration. No documented incidents of enterprise-deployed AI agents causing OT safety failures have occurred yet, but CISOs should begin mapping where agents intersect with Purdue Model levels and establish clear boundaries for agent authority over physical actuators.

Adversarial Agent Weaponisation

Adversaries are deploying agentic AI as an offensive capability, not just targeting agentic systems. The GTG-1002 campaign used jailbroken Claude Code instances to conduct largely autonomous espionage across roughly 30 organizations, with AI executing 80–90% of tactical operations at physically impossible request rates. CrowdStrike's 2026 report documented an 89% increase in AI-enabled adversary attacks, with average breakout times falling to 29 minutes. The IAPS HACCA report warns that frontier models went from near-zero to 60% success on expert-level offensive security challenges within months. Ransomware groups are integrating agentic capabilities for autonomous reconnaissance and lateral movement. International norms frameworks — including the UN GGE and OEWG processes — do not yet address autonomous AI-driven operations. CISOs should deploy behavioral detection capable of distinguishing autonomous agent activity from legitimate workflows and conduct red team exercises simulating agent-driven campaigns rather than human-paced intrusions.

Closing

The body of this report covers a wide surface: threat trends, real-world incidents, agent identity, supply chain, regulatory pressure, and a maturity model. What was a portfolio of possible threats twelve months ago is now a portfolio of documented cases. As agent autonomy increases, the line between safety failure and security failure is blurring, and deployment-layer safety belongs with the security function. Governance practices designed for static systems and quarterly cycles cannot meet incident reporting clocks measured in hours, or the pace at which new coding agents reach enterprise production.

For most organisations this is not a future problem to address. The agents are already deployed, Shadow AI is already present, and managing the growing number of non-human identities is becoming a real operational challenge. What changes is whether the governance posture is shaped intentionally, against an accurate map of the current attack surface, or shaped after the first incident.

The appendices that follow are the operational substrate for that work: the detailed agent type taxonomy, the global regulatory landscape across 42 instruments and 10 jurisdictions, ASI risk classes mapped to adoption tier, the notable projects survey, the FinBot CTF training environment, and an applied Top 10 walkthrough for personal agents. Each is intended to be readable in isolation and to support the chapters above when more depth is needed.

Appendix 1: Detailed Agent Type Taxonomy

This appendix covers the 5 agent types by operational role discussed in the Agent Taxonomy section. Each entry describes what the agent does and where it operates, including a description establishing the category's defining governance characteristics, a structured properties table, and key controls.

1.1 Enterprise Agents

Enterprise agents are AI-driven systems deployed for internal organizational use to support employee workflows, knowledge access, analytics, and automation. **The defining boundary for this category is that the agent serves internal users and operates within (but frequently reaches beyond) organizational trust boundaries.** This distinguishes enterprise agents from client-facing agents (which serve external users in adversarial environments) and from personal agents (which operate outside organizational governance entirely).

These agents commonly retrieve enterprise knowledge through RAG pipelines or direct database connections and may access sensitive information related to products, customers, operations, and strategy. Access is typically governed through RBAC or attribute-based policies. *RBAC-context discrepancies*, or the discrepancy between identity-based permissions and the contextual data retrieved by the agent, remain a persistent source of risk. Many enterprise agents incorporate function-calling capabilities that connect to external services, expanding utility but also extending the attack surface beyond internal boundaries. Logs, telemetry pipelines, and generated outputs are frequently overlooked exfiltration vectors; data that would be blocked at an API boundary may flow freely through observability infrastructure. This combination of privileged internal access and external connectivity creates the conditions for the "lethal trifecta" (access to private data, exposure to untrusted content, and the ability to communicate externally) that makes prompt injection exploitable at organizational scale. (See Threat Analysis section for detailed discussion of the lethal trifecta and Meta's complementary Agents Rule of Two.)

Autonomy Range	**Human-in-the-loop approval to autonomous multi-step workflows**
Trust Boundary Exposure	Internal systems + external APIs (web search, SaaS integrations, third-party services)

Persistence	Session-scoped to persistent memory; RAG sources dynamically updated
Primary ASI Threats	ASI01 (Goal Hijack via poisoned documents), ASI02 (Tool Misuse), ASI04 (Supply Chain via managed connectors), ASI06 (Memory & Context Poisoning), ASI03 (Identity & Privilege Abuse)
Regulatory Triggers	EU AI Act (Annex III if in covered sector), GDPR Art 22 (automated decisions), DORA (financial sector), NIS2 (critical infrastructure)
Notable Disclosures (2025–26)	EchoLeak (zero-click Copilot exfiltration), ForcedLeak (Agentforce CRM data exposure), Microsoft Copilot Studio published without authentication and frequently misconfigured agents

Key Controls: Organizations **must** validate all RAG data sources and implement output filtering to prevent exfiltration. They **should** audit data access patterns regularly, enforce strict separation between user permissions and agent capabilities, apply least-agency principles granting agents only the minimum permissions required, and review telemetry and logging pipelines to ensure sensitive data is not exposed through observability infrastructure. They **may** implement deterministic hooks at external communication boundaries to operationalize the Rule of Two, ensuring that agents with access to private data and untrusted content require human approval before communicating externally.

1.2 Coding Agents

Coding agents automate code generation, refactoring, testing, and deployment workflows. **The defining boundary for this category is direct read/write access to source code and infrastructure configurations, making these agents part of the software supply chain.** This distinguishes them from enterprise agents that may use code incidentally. Coding agents operate *on* the codebase as their primary function.

These agents integrate with source-control systems, CI/CD pipelines, and cloud APIs, giving them read/write access to sensitive repositories, deployment keys, and infrastructure. Advanced implementations execute multi-step workflows autonomously: generating code, running tests, fixing failures, committing changes, and opening pull requests. The shift toward fully autonomous operation, exemplified by the "Ralph Wiggum" [48] pattern of looped, unattended agent invocation, is accelerating faster in this category than in any other. Because coding agents operate across repositories and dependency graphs, a single compromised commit

or poisoned dependency can propagate automatically through downstream projects, amplifying the blast radius beyond the initially targeted codebase.

Containment controls designed for human-controlled execution break down when the executor is an autonomous agent. CVE-2026-22708 (Cursor allowlist bypass) and CVE-2025-59532 (Codex CLI sandbox boundary bypass) demonstrate that sandboxes and allowlists whose assumptions were calibrated for human operators become exploitable when the agent can influence its own containment environment. Organizations that treat sandboxing and allowlisting as sufficient containment for coding agents are relying on controls that the agent's architecture is positioned to undermine.

Autonomy Range	**Copilot-style (human approves each action) to fully autonomous (unattended multi-step execution)**
Trust Boundary Exposure	Source repositories, CI/CD pipelines, cloud infrastructure, package registries, developer credentials
Persistence	Session-scoped to persistent (git history as memory in autonomous patterns)
Primary ASI Threats	ASI04 (Supply Chain), ASI05 (Unexpected Code Execution / RCE), ASI02 (Tool Misuse), ASI01 (Goal Hijack via repo content)
Regulatory Triggers	NIS2 (if in critical infrastructure)
Notable Disclosures (2025–26)	CVE-2026-22708 (Cursor allowlist bypass), CVE-2025-59532 (Codex CLI sandbox boundary bypass), Replit production DB deletion, 100% of major AI coding IDEs found vulnerable in security audit (IDEsaster)[49]

Key Controls: Organizations **must** implement secret scanning on all commits, enforce least-privilege tokens, and maintain code signing and artifact verification. They **should** require mandatory human review for sensitive changes (production deployments, security-critical code, credential access), run automated security analysis before merge, apply network segmentation for agent execution environments, enforce policy-as-code gates in CI/CD pipelines, and maintain comprehensive audit logging of all agent-initiated commits and infrastructure changes. They **may** adopt multi-agent verification patterns where one agent generates code and another reviews it for security vulnerabilities and policy compliance.

1.3 Client-Facing Agents

Client-facing agents interact directly with customers, partners, or other external users. **The defining boundary for this category is public accessibility combined with transactional capability:** these agents operate in adversarial environments where any user may attempt manipulation, and they carry the heaviest regulatory burden when processing customer data in consequential decisions. This distinguishes them from enterprise agents, which serve internal users behind authentication boundaries.

Common deployments include customer service chatbots, voice assistants, onboarding workflows, and self-service portals. Unlike earlier chatbot-style systems, modern client-facing agents frequently perform operational actions: modifying account settings, initiating refunds, processing transactions, scheduling services, or triggering backend workflows. The combination of public exposure and operational authority makes them the broadest direct attack surface of any agent category. Public accessibility also makes client-facing agents the primary target for resource exhaustion attacks, where adversaries drive up compute costs or degrade service availability by flooding the agent with high-complexity requests.

Autonomy Range	Scripted responses to adaptive, action-taking agents with backend integrations
Trust Boundary Exposure	Public-facing; handles customer PII; connected to payment processors, scheduling, CRM, backend APIs
Persistence	Typically session-scoped; some maintain customer context across interactions
Primary ASI Threats	ASI01 (Goal Hijack / prompt injection from users), ASI04 (Supply Chain via managed connectors), ASI09 (Human-Agent Trust Exploitation), ASI02 (Tool Misuse via transaction abuse)
Regulatory Triggers	GDPR Art 22 (solely automated decisions with legal or similarly significant effect), CO SB 24-205 (penalties up to $20K per violation), TX TRAIGA (prohibited use boundaries), PCI DSS (if payment processing)
Notable Disclosures (2025–26)	Amazon lawsuit alleging Perplexity Comet agent accessed customer accounts without authorization

Key Controls: Organizations **must** implement bounded autonomy with explicit limits on agent actions, human escalation paths for high-impact operations, and transaction approval workflows for financial

actions. They **should** deploy runtime intent validation, rate limiting, abuse detection, input sanitization, and output filtering. They **may** implement progressive trust models where new or unverified users interact with more constrained agent capabilities.

Regulatory scope note: Texas RAIGA (HB 149) is an intent-based prohibited-uses law, not a broad high-risk framework. It applies only when AI is used for specifically restricted purposes such as facilitating self-harm, discrimination, CSAM, or rights infringement. Most standard client-facing deployments will not trigger it directly. However, because client-facing agents operate in adversarial environments where any user may attempt manipulation (ASI01), organizations must ensure that agents cannot be induced into performing prohibited activities through prompt injection or tool misuse. The compliance obligation is not to the agent's intended function but to the full range of outputs an adversary could extract from it.

1.4 Personal Agents

Personal agents run locally on user devices, outside enterprise governance structures, with the user's full system permissions. **The defining boundary for this category is the absence of an organizational control layer:** there is no RBAC, no centralized monitoring, no security review before deployment. The user is simultaneously the administrator, the operator, and the trust anchor. This distinguishes them from enterprise agents (which operate within organizational governance) and from coding agents (which, even when used individually, typically connect to enterprise-managed repositories and CI/CD systems).

This category includes self-hosted assistants (OpenClaw, Claude Desktop extensions), local coding assistants, and personal productivity agents. When employees run personal agents on work devices or connect them to work systems, the organizational attack surface expands without the CISO's knowledge or consent.

Autonomy Range	**High to fully autonomous; typically always-on with broad local permissions**
Trust Boundary Exposure	User device (full file system, shell access, messaging platforms); may connect to work systems without IT approval
Persistence	Persistent memory (SOUL.md, MEMORY.md files); skill registries; user-configured personality and behavior

Primary ASI Threats	ASI04 (Supply Chain: skill registry poisoning), ASI06 (Memory Poisoning: persistent behavioral modification), ASI10 (Rogue Agents), ASI05 (RCE via malicious skills)
Regulatory Triggers	Limited direct regulatory exposure for personal use; Shadow AI risk when used on enterprise devices/systems (triggers same obligations as enterprise agents but without controls)
Notable Disclosures (2025–26)	ClawHavoc campaign (~12% of 2,857 ClawHub skills found malicious in Feb 2026 audit), skill-embedded reverse shells, SOUL.md/MEMORY.md poisoning for time-delayed behavioral modification, Claude Desktop extension RCE

Key Controls: Organizations **must** include personal agents in shadow AI inventory programs and extend endpoint detection and response (EDR) to cover agent runtime behavior on managed devices. They **should** establish clear acceptable-use policies for personal agents on work devices, implement network-level monitoring for unexpected agent-initiated connections, and provide sanctioned alternatives to reduce shadow AI adoption. They **may** deploy application allowlisting to prevent unapproved agent runtimes on managed devices.

1.5 Infrastructure and Operations Agents

Infrastructure and operations agents manage cloud resources, CI/CD pipeline execution, monitoring, alerting, and incident response. **The defining boundary for this category is infrastructure-level permissions and blast radius:** these agents don't write code; they manage the systems code runs on. This distinguishes them from coding agents. Compromise of an infrastructure agent doesn't just expose one system; it enables lateral movement across the environment.

These agents hold cloud IAM roles, deployment credentials, and monitoring API keys, and operate across production environments.

Autonomy Range	Supervised (alert routing, dashboarding) to autonomous (auto-scaling, auto-remediation, deployment)
Trust Boundary Exposure	Cloud infrastructure, production environments, CI/CD pipelines, monitoring and alerting systems
Persistence	Typically session-scoped, but credentials and infrastructure state persist
Primary ASI Threats	ASI03 (Identity & Privilege Abuse: over-scoped infra credentials), ASI04 (Supply Chain), ASI05 (RCE via tool exploitation), ASI02 (Tool Misuse: destructive infrastructure actions)
Regulatory Triggers	NIS2 (critical infrastructure), DORA (financial sector ICT)
Notable Disclosures (2025–26)	ShadowRay 2.0 (Ray AI framework botnet), ToolShell RCE via SharePoint, ShadowMQ cluster takeover

Key Controls: Organizations **must** enforce least-privilege IAM for all agent service accounts, implement credential rotation and ephemeral token issuance, and maintain kill switches for autonomous remediation actions. They **should** segment agent execution environments from production, conduct blast-radius analysis for all agent permission grants, and monitor for anomalous infrastructure state changes. They **may** adopt infrastructure-as-code verification where agent-proposed changes are validated against policy before execution.

Appendix 2: Global Regulatory and Compliance Landscape

Disclaimer. This appendix summarizes regulatory instruments and standards relevant to agentic AI deployment. Each entry reflects the contributors' analysis of how the instrument intersects with agentic systems and is intended as a technical reference for security architects, AI engineers, and risk leaders. It is not legal advice and is not a substitute for jurisdiction-specific counsel.

This appendix covers 42 regulatory instruments, standards, and frameworks across ten jurisdictions and international standards bodies, plus 12 watchlist items tracking instruments not yet enforceable. Entries added during this revision cycle are integrated into the per-jurisdiction sections below. Each entry assesses implications for autonomous AI agents across nine dimensions: Autonomy/Tool Use, Multi-Agent Orchestration, Memory/State, Identity/Authorization, Monitoring/Auditability, Incident Reporting, Evaluation/Red-Teaming, Supply Chain, and Human Oversight. All entries verified through April 4, 2026.

2.1 European Union

Name	Impacted Verticals	General Description	Agentic AI Details
EU Artificial Intelligence Act	Healthcare, Finance, Law Enforcement, Critical Infrastructure, Public Services	Comprehensive EU regulation (2024/1689) classifying AI systems by risk tier with binding conformity assessments, transparency obligations, and enforcement penalties up to 35M euros or 7% of global turnover. GPAI Code of Practice finalized July 2025. Digital Omnibus proposal (Nov 2025) proposes to delay high-risk deadlines to December 2027, pending trilogue confirmation.	High-risk classification captures autonomous agents in Annex III sectors. Article 14 mandates human oversight with override capability. Article 72 requires post-market monitoring covering behavioral drift. Article 25 distributes supply chain liability across every component an agent invokes. GPAI systemic-risk models require red-team evaluations at RAND Security Level 3.

Council of Europe Framework Convention on AI	Public Sector, Government Services across 46 signatory nations	First binding international AI treaty (CETS No. 225). Entered into force Nov 1, 2025. Ratified by UK, France, Norway. Signed by US, EU, Israel, 10+ countries. Risk-based approach across the full AI lifecycle.	Treaty-level obligations for meaningful human control over AI systems affecting fundamental rights. Autonomous agents deployed by or on behalf of government agencies face the highest scrutiny. HUDERIA methodology endorsed as the impact assessment tool.
Digital Operational Resilience Act (DORA)	Banking, Insurance, Investment Firms, Payment Processors, Crypto-Asset Providers	Binding ICT risk management for the EU financial sector (20 entity types + critical ICT third-party providers). Enforceable Jan 17, 2025. 4-hour initial incident notification. Annual threat-led penetration testing. Fines up to 2% annual turnover.	Continuous monitoring of agent behavior in financial processes with documented baselines. Annual TLPT extends to adversarial evaluation of agent decision-making and manipulation resistance. AI model providers serving financial entities may face critical third-party provider designation with direct ESA oversight.
General Data Protection Regulation (GDPR)	Hiring, Credit Scoring, Healthcare, Finance, Insurance	Foundational EU data protection law regulating personal data processing and individual privacy rights. Fines up to 20M euros or 4% of global turnover. DPA enforcement actions increasingly target AI-driven processing.	Article 22 creates a hard floor for agent autonomy in consequential decisions by granting individuals the right to contest solely automated decisions. Data minimization and storage limitation directly constrain agent memory architectures, RAG pipelines, and cross-session context. Right to erasure applies across all memory stores an agent accesses.

NIS2 Directive	Energy, Transport, Healthcare, Water, Digital Infrastructure across the EU	Expanded cybersecurity directive for essential and important entities. 19 of 27 Member States transposed by Jan 2026. Multi-stage incident reporting: 24-hour early warning, 72-hour notification, one-month final report. Fines up to 10M euros or 2% of turnover.	Access control requirements apply to AI agents at human-equivalent rigor. 24-hour early warning timeline requires automated detection of agent-caused incidents in critical infrastructure. Article 21(2)(d) supply chain security covers all third-party tools, APIs, and model endpoints agents access at runtime.
EU Revised Product Liability Directive (2024/2853)	All sectors placing AI products or digital services on the EU market	Replaces Directive 85/374/EEC. Treats software, AI systems, and digital services as products subject to strict no-fault liability. Expanded definition of liable parties includes software developers, authorized representatives, and online platforms. Eased burden of proof for claimants in technically complex cases. No maximum liability cap.	Autonomous agents that cause harm trigger strict liability for every entity in the AI value chain, from model provider to deployer. Missing or deficient software updates create ongoing liability exposure. Agents accessing third-party tools or services extend the liability chain to those integration points. Complements the AI Act: an agent classified as high-risk under the AI Act and found defective under the RPLD exposes the deployer to both regulatory penalties and civil damages.

1. EU Artificial Intelligence Act

Key Dates: Entered into force August 1, 2024. Prohibitions effective February 2, 2025. GPAI Code of Practice finalized July 2025; provider obligations apply August 2, 2025. High-risk requirements enforceable August 2, 2026. Digital Omnibus proposal (November 2025) would delay high-risk deadlines to December 2027, pending trilogue confirmation. This is not yet finalized as of Q1 2026.

Description: The world's first comprehensive, binding AI regulatory framework. Classifies AI by risk tier: unacceptable (banned), high-risk (regulated), limited-risk (transparency), and minimal-risk. The finalized GPAI Code of Practice requires transparency documentation, copyright compliance, and for systemic-risk models, lifecycle risk management, red-team evaluations at RAND Security Level 3, and annual safety reporting. Fines reach 35 million euros or 7% of global turnover for prohibited practices.

Agentic Implications:

- Autonomy/Tool Use (High): Article 6 and Annex III classify autonomous safety-critical AI as high-risk by default; each tool integration point becomes a conformity assessment boundary.
- Monitoring/Auditability (High): Article 72 mandates post-market monitoring covering behavioral drift, not just static performance; Article 62 requires mandatory serious incident reporting.
- Supply Chain (High): Article 25 distributes responsibilities along the AI value chain; deployers must verify provider compliance across every component an agent invokes.
- Human Oversight (High): Article 14 mandates qualified personnel who can override autonomous decisions, requiring strategic control points and risk-based escalation rather than per-action approval.
- Evaluation/Red-Teaming (High): GPAI systemic-risk models require red-team evaluations at RAND Security Level 3, lifecycle risk management, and annual safety reporting.

2. Council of Europe Framework Convention on AI

Key Dates: Opened September 5, 2024 (CETS No. 225). Entered into force November 1, 2025. Ratified by UK, France, Norway. Signed by US, EU, Israel, and 10+ countries.

Description: First binding international AI treaty. Covers full AI lifecycle with risk-based approach. Applies to public authorities and private actors on their behalf. HUDERIA methodology endorsed as the assessment tool.

Agentic Implications:

- Human Oversight (High): Treaty-level obligations for meaningful human control over AI systems affecting fundamental rights. Autonomous agents deployed by or on behalf of government agencies face the highest scrutiny.
- Monitoring/Auditability (Medium): Requires appropriate measures to identify, assess, and mitigate risks, implying ongoing evaluation for adaptive systems.
- Identity/Authorization (Medium): Accountability requirements demand clear attribution of agent actions to responsible public-sector entities.

3. Digital Operational Resilience Act (DORA)

Key Dates: Full application January 17, 2025. First registers of information due April 30, 2025. ESAs designate critical ICT third-party providers November 2025.

Description: Binding ICT risk management for the EU financial sector covering 20 entity types plus critical ICT third-party providers. Five pillars: ICT risk management, incident reporting (4-hour initial notification), resilience testing, third-party risk, and information sharing. Fines up to 2% annual turnover.

Agentic Implications:

- Monitoring/Auditability (High): Continuous monitoring of agent behavior in financial processes with documented baselines and automated deviation detection.
- Evaluation/Red-Teaming (High): Annual TLPT requirements extend to adversarial evaluation of agent decision-making, tool invocation patterns, and manipulation resistance.
- Supply Chain (High): AI model and orchestration platform vendors may face critical third-party provider designation with direct ESA oversight.
- Incident Reporting (High): 4-hour initial incident notification after classification as major, with classification itself due within 24 hours of awareness. Agents operating in financial processes must support automated incident detection at this tempo.

4. General Data Protection Regulation (GDPR)

Key Dates: Enforceable May 25, 2018. Ongoing DPA enforcement actions targeting AI-driven processing.

Description: Foundational data protection framework governing AI data processing across the EEA. Article 22 protects against solely automated decisions with legal or significant effects. Fines up to 20 million euros or 4% of global turnover.

Agentic Implications:

- Memory/State (High): Data minimization, purpose limitation, and storage limitation directly constrain agent memory architectures; persistent stores, RAG pipelines, and cross-session context require automated lifecycle management and right-to-erasure compliance.
- Human Oversight (High): Article 22 creates a hard floor for agent autonomy in consequential decisions.
- Identity/Authorization (Medium): Data subject rights require agents to accurately identify and respond to individual data requests across all processing contexts, including context assembled dynamically at runtime.

5. NIS2 Directive

Key Dates: In force January 16, 2023. Transposition deadline October 2024; 19 of 27 Member States transposed by January 2026. ENISA Cyber Stress Testing Handbook published May 2025.

Description: Expands cybersecurity requirements for essential and important entities in critical sectors. Mandatory multi-stage incident reporting: 24-hour early warning, 72-hour notification, one-month final report. Fines up to 10 million euros or 2% of turnover for essential entities.

Agentic Implications:

- Identity/Authorization (High): Access control requirements apply to AI agents at human-equivalent rigor.
- Incident Reporting (High): 24-hour early warning timeline requires automated detection of agent-caused incidents.
- Supply Chain (High): Article 21(2)(d) supply chain security covers all third-party tools and services agents access at runtime.
- Monitoring/Auditability (High): ENISA Cyber Stress Testing Handbook provides supervisors with a live-fire drill methodology for critical sectors, applicable to agent-dependent infrastructure.

6. EU Revised Product Liability Directive (2024/2853)

Key Dates: Adopted October 23, 2024. Published in Official Journal November 18, 2024. Entered into force December 9, 2024. Member States must transpose into national law by December 9, 2026. Directive 85/374/EEC repealed from December 9, 2026. AI Liability Directive proposal withdrawn by the European Commission in February 2025.

Description: Replaces the 1985 Product Liability Directive with a modernized strict no-fault liability regime covering digital-age products. Explicitly includes software (embedded and standalone), AI systems, and digital services within the definition of "product." Expands the range of liable parties to include software developers, authorized representatives, fulfillment service providers, and, under specified conditions, online platforms and distributors. Eases the burden of proof for claimants in technically complex cases through rebuttable presumptions and mandatory evidence disclosure. Removes previous maximum liability caps. Extends the long-stop period to 25 years for latent personal injury. Following the withdrawal of the proposed AI Liability Directive in February 2025, this Directive is the primary civil liability mechanism for AI-caused harm in the EU.

Agentic Implications:

- Supply Chain (High): Every entity in the agentic AI value chain, from foundation model provider through orchestration framework vendor to deploying organization, is a potential defendant. Tool providers, MCP server operators, and plugin developers whose components an agent invokes at runtime fall within the expanded definition of liable parties.
- Monitoring/Auditability (High): The Directive's evidence disclosure provisions mean organizations must maintain records sufficient to demonstrate product safety. For agents that compose behavior at runtime, this requires trajectory-level logging of tool invocations, state changes, and decision rationale.

- Autonomy/Tool Use (High): Failure to provide adequate software updates or cybersecurity protections creates ongoing liability exposure. Agents that self-modify or load new capabilities post-deployment trigger continuing obligations for the entity that placed the system on the market.
- Human Oversight (Medium): While the Directive does not prescribe specific oversight mechanisms, the strict liability standard incentivizes deployers to implement human review at high-consequence decision points as a risk mitigation measure.
- Evaluation/Red-Teaming (Medium): The rebuttable presumption of defectiveness in technically complex cases creates a strong incentive for pre-deployment and ongoing adversarial testing, as the burden shifts to the defendant to prove the product was not defective.

2.2 United States: Federal

Name	Impacted Verticals	General Description	Agentic AI Details
DHS AI Safety and Security Guidelines	Energy, Transportation, Water, Healthcare, Telecom	Security guidelines for AI in critical infrastructure aligning with US federal cybersecurity frameworks. Developed in response to EO 14110. Integrates NIST AI RMF and addresses AI-driven cybersecurity threats, operational risks, and supply chain vulnerabilities across 16 critical infrastructure sectors.	Framework for secure agent deployment in critical infrastructure. Agents operating across CI sectors must meet sector-specific cybersecurity requirements and resilience standards. Continuous risk assessment model for AI applications. Mandatory sector-wide reporting protocols for AI-related incidents.
EO 14365: Ensuring a National Policy Framework for AI	All sectors deploying AI across US states	Signed Dec 11, 2025. Directs federal preemption of state AI laws. DOJ AI Litigation Task Force established. Commerce state law	Creates a volatile compliance environment for multi-state agentic deployments. Directly threatens state incident reporting mandates (CA 15-day, NY 72-hour). Organizations

		report due March 11, 2026. Explicitly named Colorado AI Act.	should plan for the most restrictive timeline until preemption is resolved.
EO 14179: Removing Barriers to American Leadership in AI	Federal Infrastructure, Defense, Energy, AI Development	Signed Jan 20, 2025. Revoked Biden EO 14110, eliminating mandatory safety testing, reporting thresholds, and the dual-use foundation model classification system. Shifted to industry self-regulation.	No federal obligation for pre-deployment safety testing of frontier models used in agentic systems. Limited federal visibility into agent deployment scale, capabilities, and incidents. Burden shifts to voluntary frameworks and state laws.
NIST AI 600-1 (Generative AI Profile)	Federal Contractors, Regulated Industries, All Sectors (voluntary)	Companion resource to AI RMF 1.0 for generative AI. Published July 26, 2024. Identifies 12 GAI-specific risk categories. Sector-agnostic. Developed pursuant to EO 14110 with input from the GAI Public Working Group.	Covers risks directly applicable to foundation models underlying agentic systems: confabulation, prompt injection, data poisoning, data privacy leakage, information integrity, and CBRN information access. Suggested actions map to AI RMF Govern/Map/Measure/Manage functions with GAI-specific controls.
NIST AI Risk Management Framework 1.0	Federal Contractors, Regulated Industries, All Sectors (voluntary)	Voluntary four-function framework: Govern, Map, Measure, Manage. Released Jan 2023. De facto US federal AI risk management standard. Referenced as safe harbor in Texas RAIGA.	MAP requires documenting full scope of agent tools and autonomy boundaries. GOVERN addresses oversight roles. MANAGE includes human-in-the-loop and incident response planning. Extended by Cyber AI Profile (IR 8596) for agent-specific security controls.

NIST Cyber AI Profile (IR 8596)	Federal Agencies, Critical Infrastructure, AI Developers	First NIST publication mapping AI cybersecurity controls to CSF 2.0. Provides foundational coverage but explicitly does not address multi-agent architectures, agent identity, or inter-agent trust; OWASP Agentic Top 10 fills this gap directly. Final publication status unknown at time of print; preliminary draft comment period closed January 30, 2026.	Mandates unique agent identities, defined permissions, continuous verification, and least-privilege access. Explicitly restricts agent code execution capabilities. Adversarial training and adaptive learning scenarios as testing categories. Strongest federal guidance on agent identity to date.
TAKE IT DOWN Act	Social Media, Content Platforms, AI Content Generation	First US federal law on AI-generated harmful content. Signed May 2025. Criminalizes non-consensual intimate digital depictions including deepfakes. 48-hour platform removal requirement.	Content-generation agents face criminal liability for producing prohibited outputs. First US federal criminal prohibition directly applicable to AI-generated content. Requires output filtering at every agent generation point.

1. DHS AI Safety and Security Guidelines

Key Dates: Published April 2024 in response to Executive Order 14110. Developed by DHS in coordination with CISA, with input from 16 Sector Risk Management Agencies. CISA/NCSC co-developed Guidelines for Secure AI System Development published November 2023. Cross-Sector AI Risk Analysis completed January 2024. CISA/NSA joint guidance on AI agents in OT published December 2025, co-authored with eight allied nations.

Description: Risk-based framework for securing AI systems across 16 critical infrastructure sectors. Integrates the NIST AI RMF and addresses AI-driven cybersecurity threats, operational risks, and supply chain vulnerabilities. Five focus areas: AI risk management for critical infrastructure, cybersecurity controls for AI systems, governance and compliance alignment, AI supply chain security and procurement standards, and incident response and AI system resilience. The December 2025 CISA/NSA joint guidance extends coverage to AI agents operating in OT environments, establishing four core principles for secure integration in industrial control systems. Note: the original guidelines were developed pursuant to EO 14110, which has since been revoked. DHS has indicated it will continue to update the guidelines, and the recommendations

align with established cybersecurity best practices that remain operative independent of the originating executive order.

Agentic Implications:

- Monitoring/Auditability (High): Continuous risk assessment model for AI applications across critical infrastructure sectors. Mandates sector-wide reporting protocols for AI-related incidents, including adversarial attacks, system failures, and unanticipated AI behaviors.
- Supply Chain (High): Establishes security guidelines for AI components, training data integrity, and AI supply chain risk management. Minimum security requirements for AI vendors and third-party developers working with critical infrastructure operators.
- Autonomy/Tool Use (High): December 2025 OT guidance establishes clear boundaries for agent authority over physical actuators in industrial environments. Agents operating in OT must respect Purdue Model layer separation and safety-instrumented system constraints under IEC 62443.
- Incident Reporting (Medium): Defines AI-specific recovery and continuity planning strategies. Sector-wide reporting covers agent-caused anomalies, not just traditional cybersecurity incidents.
- Evaluation/Red-Teaming (Medium): Aligns with NIST adversarial machine-learning taxonomy for testing AI systems in critical infrastructure contexts.

2. Executive Order 14365: Ensuring a National Policy Framework for AI

Key Dates: December 11, 2025. DOJ AI Litigation Task Force established by January 10, 2026. Commerce state law report, FTC policy statement, and FCC rulemaking all due March 11, 2026.

Description: Pushes federal preemption of state AI laws. DOJ task force to challenge state laws through litigation. Commerce to evaluate state AI regulations. FTC and FCC directed to develop federal alternatives. Explicitly names the Colorado AI Act.

Agentic Implications:

- Monitoring/Auditability (High): Creates a volatile compliance environment for multi-state agentic deployments. Organizations cannot yet determine which state obligations will survive federal preemption.
- Incident Reporting (Contested): Directly threatens state-level reporting mandates (CA 15-day, NY 72-hour). Organizations should plan for the most restrictive timeline until preemption is resolved.
- Human Oversight (Contested): State mandates for meaningful human review of AI decisions face potential invalidation. Organizations should maintain oversight controls regardless, as federal alternatives are expected to impose comparable requirements.

3. Executive Order 14179: Removing Barriers to American Leadership in AI

Key Dates: January 20, 2025. Revoked Biden EO 14110 immediately.

Description: Eliminated mandatory safety testing, reporting thresholds, and the dual-use foundation model classification system. Shifted from prescriptive safety mandates to industry self-regulation.

Agentic Implications:

- Evaluation/Red-Teaming (Reduced): No federal obligation for pre-deployment red-team evaluations of frontier models used in agentic systems.
- Monitoring/Auditability (Reduced): Limited federal visibility into frontier model deployments, capabilities, and incident histories. Burden shifts to voluntary frameworks and state laws.

4. NIST AI 600-1 (Generative AI Profile)

Key Dates: Initial public draft April 29, 2024. Final publication July 26, 2024. Developed pursuant to Section 4.1(a)(i)(A) of EO 14110 with input from the Generative AI Public Working Group.

Description: Cross-sectoral companion resource to the AI RMF 1.0 for generative AI. Identifies 12 risk categories unique to or exacerbated by GAI: CBRN information, confabulation, data privacy, environmental impacts, harmful bias and homogenization, human-AI configuration, information integrity, information security, intellectual property, obscene/degrading content, and value chain/component integration. Provides suggested actions organized by the AI RMF's Govern, Map, Measure, and Manage functions. Sector-agnostic and voluntary. Note: developed under EO 14110, which has since been revoked, but the publication remains available and applicable as voluntary guidance.

Agentic Implications:

- Evaluation/Red-Teaming (High): Suggested actions include structured pre-deployment testing for confabulation, prompt injection resilience, and adversarial robustness. GAI-specific red-teaming guidance covers scenarios directly relevant to foundation models underlying agentic systems.
- Monitoring/Auditability (High): Incident disclosure considerations address post-deployment monitoring of generative outputs. Provenance tracking for GAI-generated content applicable to agent output chains.
- Memory/State (Medium): Data privacy risk category addresses training data memorization, inference-time data leakage, and the ability of models to correctly infer PII not present in training data. These risks compound in agents with persistent memory and cross-session context.
- Supply Chain (Medium): Value chain and component integration risk category covers third-party model integration, fine-tuning risks, and downstream deployment scenarios. Applicable to multi-model agentic architectures.
- Human Oversight (Medium): Human-AI configuration risk category addresses over-reliance, automation bias, and inappropriate anthropomorphization. Directly relevant to human-agent trust dynamics in agentic deployments.

5. NIST AI Risk Management Framework 1.0

Key Dates: Released January 26, 2023. Extended by Cyber AI Profile (IR 8596) December 2025. Referenced as safe harbor in Texas RAIGA.

Description: Voluntary framework with four functions: Govern, Map, Measure, Manage. De facto US federal standard for AI risk management.

Agentic Implications:

- Autonomy/Tool Use (High): MAP requires documenting full scope of agent tools, decision-making boundaries, and autonomy conditions.
- Human Oversight (High): GOVERN addresses oversight roles; MANAGE includes human-in-the-loop requirements demanding clear escalation thresholds and authority boundaries.
- Monitoring/Auditability (Medium): MEASURE function requires metrics for trustworthiness characteristics. For agentic systems, this extends to behavioral drift detection, goal alignment verification, and decision-quality assessment.

6. NIST Cyber AI Profile (IR 8596)

Key Dates: Preliminary draft December 16, 2025. Comment closed January 30, 2026. Final expected H2 2026.

Description: First NIST document explicitly addressing agentic AI security. Maps AI cybersecurity to CSF 2.0 across three focus areas: Secure AI systems, Defend with AI, Thwart AI-enabled attacks. Developed with 6,500+ contributors. Provides foundational coverage but explicitly does not address multi-agent architectures, agent identity, or inter-agent trust; OWASP Agentic Top 10 fills this gap directly.

Agentic Implications:

- Identity/Authorization (High): Mandates unique agent identities, defined permissions, continuous verification, and least-privilege access. Strongest federal guidance on agent identity to date.
- Autonomy/Tool Use (High): Explicitly restricts agent code execution capabilities, treating agents as bounded actors with defined operational envelopes.
- Evaluation/Red-Teaming (High): Adversarial training and adaptive learning scenarios as explicit testing categories. Covers manipulation resistance and behavioral consistency under adversarial conditions.
- Monitoring/Auditability (Medium): Maps to CSF 2.0 Detect and Respond functions with AI-specific indicators. Continuous verification requirements imply runtime behavioral monitoring.

7. TAKE IT DOWN Act

Key Dates: Signed May 19, 2025. Platform compliance within 12 months.

Description: First US federal law on AI-generated harmful content. Criminalizes non-consensual intimate digital depictions including deepfakes. Requires 48-hour platform removal.

Agentic Implications:

- Autonomy/Tool Use (Medium): Content-generation agents face criminal liability for producing prohibited outputs. First US federal criminal prohibition applicable to AI-generated content. Requires output filtering at every agent generation point.
- Monitoring/Auditability (Medium): 48-hour removal requirement demands automated detection of prohibited AI-generated content. Platforms hosting agent-generated outputs must maintain monitoring capabilities.

2.3 United States: State

Over 145 state AI laws enacted in 2025. This section covers the most consequential for agentic deployments.

Name	Impacted Verticals	General Description	Agentic AI Details
California FEHA Regulations	Employment in California	Employer liability for vendor-provided AI in employment decisions. Effective October 2025.	Employers liable for discriminatory outcomes from vendor-supplied agentic hiring tools regardless of vendor representations.
California SB 53 (Transparency in Frontier AI Act)	Frontier AI Developers (>10^26 FLOPs or >$500M revenue)	Signed Sep 29, 2025. Effective Jan 1, 2026. Requires transparency reports, safety frameworks, 15-day incident reporting, whistleblower protections. Federal deference provision accepts EU AI Act compliance. Penalties up to $1M per violation.	Transparency reports must document safety testing of agentic use cases. Covers fine-tuned derivatives. 15-day reporting for critical safety incidents involving unauthorized agent actions or harmful outputs.

Colorado SB 24-205 (Consumer Protections for AI)	Employment, Healthcare, Finance, Education, Legal, Housing in Colorado	First comprehensive US state high-risk AI law. Delayed to June 30, 2026 by SB 25B-004. AG-only enforcement at $20K per violation per consumer. 60-day cure period. Named in Dec 2025 Trump EO as preemption target.	Broad consequential-decision definition captures most agentic deployments in covered sectors. Per-consumer penalties accumulate rapidly for agents making thousands of daily automated decisions. Mandatory meaningful human review of adverse decisions. Faces highest federal preemption risk.
Illinois HB 3773	Employment (Hiring, Promotion, Termination) in Illinois	Effective Jan 1, 2026. Prohibits employment discrimination through AI. Private right of action. ZIP code proxy prohibition.	Private right of action creates direct liability exposure for agentic hiring systems, unlike AG-only enforcement in Colorado.
New Jersey N.J.A.C. 13:16	Employment in New Jersey	Pre-deployment bias testing for automated employment tools. Effective December 2025.	Automated hiring agents require third-party bias audits before deployment. Testing must cover protected class impact across agent decision outputs.
New York RAISE Act	Frontier AI Developers (>10^26 FLOPs or >$100M compute)	Most stringent US state AI safety law. Signed Dec 19, 2025. Effective Jan 1, 2027. Creates DFS oversight office. Penalties: $1M first violation, $3M repeat. Knowledge distillation provision extends coverage.	72-hour incident reporting (strictest in US) requires automated detection and classification. Testing must have sufficient detail for third-party replication. Distillation provision closes the derivative model coverage gap for compressed agentic deployments.

Texas RAIGA (HB 149)	Employment, Education, Public Services in Texas	Effective Sep 1, 2025. Prohibits AI for restricted purposes (self-harm, discrimination, CSAM, rights infringement). NIST AI RMF compliance provides affirmative defense.	Hard boundaries on agent capabilities regardless of architecture. Agents must resist manipulation into prohibited activities through injection or tool misuse. NIST safe harbor makes framework compliance commercially significant.

1. California FEHA Regulations

Key Dates: Effective October 2025.

Description: Extends employer liability under the Fair Employment and Housing Act to cover discriminatory outcomes from vendor-provided AI tools used in employment decisions. Employers cannot disclaim responsibility by citing vendor assurances.

Agentic Implications:

- Human Oversight (High): Employers bear full liability for discriminatory outcomes from agentic hiring tools regardless of vendor representations. Creates direct incentive for meaningful pre-deployment testing and ongoing monitoring of vendor-supplied agent behavior.
- Evaluation/Red-Teaming (Medium): Practical compliance requires employers to independently validate vendor AI systems for disparate impact, not rely on vendor-provided audit reports alone.

2. California SB 53 (Transparency in Frontier AI Act)

Key Dates: Signed September 29, 2025. Effective January 1, 2026.

Description: Applies to frontier models (>10^26 FLOPs) and large developers (>$500M revenue). Requires transparency reports, safety frameworks, 15-day incident reporting, whistleblower protections. Federal deference provision accepts EU AI Act compliance. Penalties up to $1M per violation.

Agentic Implications:

- Evaluation/Red-Teaming (High): Transparency reports must document safety testing of agentic use cases; covers fine-tuned derivatives.
- Incident Reporting (High): 15-day window for critical safety incidents involving unauthorized agent actions or harmful outputs.

- Supply Chain (Medium): Coverage of fine-tuned derivatives means organizations deploying custom agents built on frontier models inherit the developer's reporting and transparency obligations.

3. Colorado SB 24-205 (Consumer Protections for AI)

Key Dates: Signed May 2024. Delayed to June 30, 2026 by SB 25B-004. Explicitly named in December 2025 Trump EO.

Description: First comprehensive US state high-risk AI law. Covers consequential decisions in employment, housing, healthcare, education, finance, and legal services. AG-only enforcement at $20,000 per violation per consumer. 60-day cure period.

Agentic Implications:

- Autonomy/Tool Use (High): Broad consequential-decision definition captures most agentic deployments in covered sectors; per-consumer penalties accumulate rapidly for agents making thousands of daily decisions.
- Human Oversight (High): Mandatory meaningful human review of adverse decisions. The scale mismatch between agent decision velocity and human review capacity makes this the most operationally challenging requirement in the law.
- Monitoring/Auditability (Medium): Annual impact assessments and documented risk management practices. Faces highest federal preemption risk among state AI laws.

4. Illinois HB 3773

Key Dates: Effective January 1, 2026.

Description: Prohibits employment discrimination through AI systems. Private right of action (unlike AG-only enforcement in Colorado). ZIP code proxy prohibition prevents using geographic data as a stand-in for protected characteristics.

Agentic Implications:

- Human Oversight (High): Private right of action creates direct liability exposure for agentic hiring systems, exposing deployers to individual lawsuits rather than solely regulatory enforcement.
- Evaluation/Red-Teaming (Medium): ZIP code proxy prohibition requires testing agent decision logic for indirect discrimination pathways that standard bias audits may miss.

5. New Jersey N.J.A.C. 13:16

Key Dates: Effective December 2025.

Description: Requires pre-deployment bias testing for automated employment decision tools. Applies to AI systems used in hiring, promotion, and termination decisions within New Jersey.

Agentic Implications:

- Evaluation/Red-Teaming (High): Automated hiring agents require third-party bias audits before deployment. Testing must cover protected class impact across agent decision outputs.
- Monitoring/Auditability (Medium): Pre-deployment testing requirement means agents that self-modify or retrain post-deployment may trigger re-testing obligations, though the regulation does not explicitly address continuous learning systems.

6. New York RAISE Act

Key Dates: Signed December 19, 2025. Effective January 1, 2027.

Description: Most stringent US state AI safety law. Covers frontier models (>10^26 FLOPs or >$100M compute). 72-hour incident reporting (strictest in US). Creates DFS oversight office. Knowledge distillation provision extends coverage to compressed model derivatives. Penalties: $1M first, $3M repeat.

Agentic Implications:

- Incident Reporting (High): 72-hour window requires automated incident detection and classification. Strictest state-level reporting timeline in the US.
- Evaluation/Red-Teaming (High): Testing must have sufficient detail for third-party replication. This standard exceeds typical internal red-team documentation practices.
- Supply Chain (Medium): Distillation provision closes the derivative model gap. Compressed or distilled agentic models inherit the full obligations of their parent frontier model.

7. Texas RAIGA (HB 149)

Key Dates: Effective September 1, 2025.

Description: Prohibits AI for restricted purposes (self-harm, discrimination, CSAM, rights infringement). NIST AI RMF compliance provides affirmative defense. AG-only enforcement. Civil penalties: $10,000 to $12,000 per curable violation, up to $200,000 per uncurable violation. 60-day cure window.

Agentic Implications:

- Autonomy/Tool Use (High): Hard boundaries on agent capabilities regardless of architecture. Agents must resist manipulation into prohibited activities through injection or tool misuse. The compliance obligation extends to the full range of outputs an adversary could extract, not just intended function.
- Evaluation/Red-Teaming (Medium): NIST AI RMF safe harbor makes framework compliance commercially significant. Organizations deploying agentic systems in Texas gain affirmative defense through documented NIST alignment.

2.4 United Kingdom

Name	Impacted Verticals	General Description	Agentic AI Details
UK AI Cyber Security Code of Practice	Telecom, Critical Infrastructure, AI Developers/Deployers	Voluntary 13-principle code published Jan 31, 2025. Addresses prompt injection, data poisoning, model obfuscation. Submitted to ETSI for international baseline standards.	Prompt injection and data poisoning controls apply to every agent tool integration point. Data integrity protections extend to RAG knowledge bases and agent memory stores. Supply chain provenance required for every component in the agent execution chain.
UK AI Regulatory Framework	All sectors via sector-specific regulators (FCA, ICO, CMA, Ofcom)	Principles-based, distributed across existing regulators. Five principles: safety, transparency, fairness, accountability, contestability. No AI Bill expected before late 2026. AI Safety Institute conducts frontier evaluations.	Multi-sector agentic platforms face different requirements per regulator. Expected ICO statutory code on automated decision-making will create binding rights to challenge agent decisions affecting UK residents.

1. UK AI Cyber Security Code of Practice

Key Dates: Published January 31, 2025. Submitted to ETSI for international baseline standards. Informed ETSI EN 304 223.

Description: Voluntary 13-principle code addressing AI-specific threats: prompt injection, data poisoning, model obfuscation, adversarial manipulation. Lifecycle coverage from design through end-of-life. Positioned as the UK's primary AI-specific cybersecurity guidance pending legislation.

Agentic Implications:

- Autonomy/Tool Use (High): Prompt injection and data poisoning controls apply to every agent tool integration point. Principles explicitly address agentic attack vectors.

- Memory/State (High): Data integrity protections extend to RAG knowledge bases, conversation histories, and persistent agent memory stores.
- Supply Chain (High): Third-party model and data provenance required for every component in the agent execution chain. Covers the full dependency graph including dynamically discovered tools.
- Evaluation/Red-Teaming (Medium): Adversarial testing principles align with agent-specific threat models. Informs ETSI EN 304 223 baseline standard, positioning UK guidance for international adoption.

2. UK AI Regulatory Framework

Key Dates: White Paper March 2023. ICO statutory code expected autumn 2025. Data Use and Access Act AI assessment due March 2026. No AI Bill before late 2026.

Description: Principles-based, distributed across sector regulators (FCA, ICO, CMA, Ofcom). Five principles: safety/robustness, transparency, fairness, accountability, contestability. Prioritizes AI Growth Zones and AI Safety Institute over legislation. AI Safety Institute conducts frontier model evaluations including agentic capability assessments.

Agentic Implications:

- Human Oversight (High): Contestability principle combined with expected ICO provisions will create binding rights to challenge agent decisions affecting UK residents. Multi-sector agents face different contestability requirements per regulator.
- Monitoring/Auditability (Medium): AI Safety Institute frontier evaluations include assessment of autonomous capabilities and self-replication risk (RepliBench). Results inform regulatory posture across sectors.
- Identity/Authorization (Medium): FCA and ICO may impose sector-specific identity and access requirements for agents operating in financial services and data processing respectively.

2.5 Asia-Pacific and Middle East

Name	Impacted Verticals	General Description	Agentic AI Details
China AI Governance Framework	Finance, Healthcare, Public Services, Content Platforms within China	Binding sector-specific regulations: GenAI Interim Measures (Jul 2023), AI Labeling Measures (Sep 2025), CSL amendments adding AI provisions (2025). Algorithmic registration with CAC required.	Embedded provenance metadata (GB45438-2025) in all agent outputs exceeds most jurisdictions' transparency requirements. Real-time content compliance checks required at every agent output point. Data localization requirements constrain cross-border agent architectures.
India AI Governance Guidelines	Public Services, Financial Services, Digital Platforms, Multi-Sector AI Deployments	Grounded in seven principles. Emphasizes calibrated oversight through infrastructure enablement, capacity building, regulatory gap reviews, graded liability, and institutional mechanisms. Released November 5, 2025 by MeitY under the IndiaAI Mission.	Enables autonomy within bounded risk frameworks. High-risk agentic deployments expected to operate through sandbox pilots and techno-legal safeguards. Strong human-centric oversight under the People-first principle with defined accountability allocation.
Japan AI Promotion Act	Public Sector, Private AI Deployment	Enacted 2025. AI Basic Plan published Dec 2025. Soft-law approach with AI Strategy Headquarters and principles-based governance.	Minimal binding requirements for agentic deployments. Sector-specific guidelines expected through existing regulators. Track financial, healthcare, and manufacturing regulator guidance.

SDAIA Ethics Principles (Saudi Arabia)	All Sectors	National framework for responsible AI deployment emphasizing fairness, transparency, accountability, privacy, and human oversight throughout the AI lifecycle. Published 2023.	Voluntary principles apply to autonomous and agentic AI. Emphasize ethical design, transparency, and human oversight. No binding enforcement mechanism.
Singapore Model Governance Framework for Agentic AI	All Sectors (Voluntary)	World's first agentic AI governance framework. Announced Jan 22, 2026 at WEF Davos. Developed by IMDA. Four governance dimensions. Open for public feedback.	Explicitly addresses agent action-space and recommends limiting tools to minimum necessary. Only framework globally requiring testing of both individual and multi-agent interactions. Requires agents to log plans for evaluation. Mandatory kill-switch capability.
South Korea AI Basic Act	Healthcare, Energy, Public Services, Credit, High-Performance AI	First comprehensive AI law in Asia-Pacific. Effective Jan 22, 2026. Consolidates 19 bills. Extraterritorial. High-performance threshold at 10^26 FLOPs.	Broad high-impact classification captures agentic systems in covered sectors. Mandatory oversight and explainability for high-impact AI. Mandatory incident reporting for high-performance AI. Foreign companies must designate a domestic representative.
UAE Charter for AI (2024)	All Sectors	National guideline with 12 ethical principles for transparent, safe, and inclusive AI use, aligned with UAE values and global governance practices.	Applies to autonomous agents. Encourages human oversight, safety, explainability, and value alignment. Voluntary, non-binding.

1. China AI Governance Framework

Key Dates: GenAI Interim Measures effective July 2023. AI Labeling Measures effective September 2025. CSL amendments adding AI provisions 2025. Algorithmic registration with CAC required.

Description: Binding sector-specific regulations combined with content controls. Mandatory AI content labeling with embedded provenance metadata (GB45438-2025). CSL amendments add AI risk assessment requirements. Algorithmic registration with the Cyberspace Administration of China required for recommendation, deep synthesis, and generative AI services.

Agentic Implications:

- Monitoring/Auditability (High): Embedded provenance metadata in all agent outputs exceeds most jurisdictions' transparency requirements. Algorithmic registration creates a government-maintained inventory of deployed AI systems.
- Autonomy/Tool Use (High): Real-time content compliance checks required at every agent output point. Agents must filter outputs against prohibited content categories before delivery.
- Memory/State (Medium): Data localization requirements constrain cross-border agent architectures. Training data, inference data, and agent memory stores for Chinese users must remain within China's borders.

2. India AI Governance Guidelines

Key Dates: Released November 5, 2025 by Ministry of Electronics and Information Technology (MeitY) under the IndiaAI Mission.

Description: Principles-based, agile regulatory model grounded in seven core principles: Trust, People-first, Innovation over restraint, Fairness and equity, Accountability, Understandable by design, and Safety, resilience, and sustainability. Adapted from the Reserve Bank of India's FREE-AI approach. Emphasizes calibrated governance through structured infrastructure access, national capacity building, iterative policy development, India-specific risk mitigation, graded liability models, and institutional oversight bodies including the AI Governance Group, Technology Policy Expert Committee, and AI Safety Institute.

Agentic Implications:

- Human Oversight (High): People-first principle places strong emphasis on human-centric design, requiring meaningful human oversight and clear allocation of responsibility for agent actions.
- Evaluation/Red-Teaming (Medium): AISI-led evaluation mechanisms reinforce logging, bias mitigation, and incident response practices relevant to multi-agent and self-evolving architectures.
- Monitoring/Auditability (Medium): Transparency and "understandable by design" requirements elevate explainability and reporting obligations. High-risk agentic deployments expected to operate through sandbox pilots with defined techno-legal safeguards.

3. Japan AI Promotion Act

Key Dates: Enacted 2025. AI Basic Plan published December 2025.

Description: Soft-law approach with AI Strategy Headquarters and principles-based governance. Minimal binding requirements; sector-specific guidelines expected through existing regulators.

Agentic Implications:

- Human Oversight (Low): Principles-based approach encourages but does not mandate specific oversight mechanisms for autonomous systems.
- Monitoring/Auditability (Low): No binding monitoring or reporting requirements. Track financial, healthcare, and manufacturing regulator guidance for sector-specific obligations as they emerge.

4. SDAIA Ethics Principles (Saudi Arabia)

Key Dates: Published 2023 by the Saudi Data and Artificial Intelligence Authority.

Description: National framework for responsible AI deployment emphasizing fairness, transparency, accountability, privacy, and human oversight throughout the AI lifecycle.

Agentic Implications:

- Human Oversight (Medium): Voluntary principles emphasize human oversight and ethical design for autonomous systems. No binding enforcement mechanism or penalty regime.
- Monitoring/Auditability (Low): Principles encourage transparency and accountability but establish no specific monitoring requirements or compliance pathway for agentic deployments.

5. Singapore Model Governance Framework for Agentic AI

Key Dates: Announced January 22, 2026 at WEF Davos. Developed by IMDA. Open for public feedback.

Description: World's first governance framework specifically designed for agentic AI. Four dimensions: (1) Assess and bound risks upfront, (2) Human accountability with defined checkpoints, (3) Technical controls across lifecycle, (4) End-user responsibility.

Agentic Implications:

- Autonomy/Tool Use (High): Explicitly addresses agent action-space; recommends limiting tools to minimum necessary. Only framework globally that specifically constrains the set of capabilities available to an agent at runtime.
- Multi-Agent Orchestration (High): Only framework globally requiring testing of both individual and multi-agent interactions. Addresses coordination risks, cascading failures, and emergent behaviors in multi-agent compositions.

- Memory/State (High): Requires agents to log plans for evaluation; persistent memory actively managed with defined retention and review policies.
- Human Oversight (High): Defined checkpoints, automation bias mitigation, mandatory kill-switch capability. Positions human accountability as a structural requirement rather than an aspirational principle.

6. South Korea AI Basic Act

Key Dates: Enacted January 21, 2025. Effective January 22, 2026. Enforcement decrees under development.

Description: First comprehensive AI law in Asia-Pacific, consolidating 19 bills. Risk-based classification for high-impact AI (healthcare, energy, public services, credit). Extraterritorial application. High-performance threshold at 10^26 FLOPs. Foreign companies must designate a domestic representative.

Agentic Implications:

- Autonomy/Tool Use (High): Broad high-impact classification captures agentic systems in covered sectors. High-performance threshold aligns with frontier model definitions in US state laws.
- Human Oversight (High): Mandatory oversight and explainability for high-impact AI. Extends to autonomous decision-making systems operating in covered domains.
- Incident Reporting (High): Mandatory incident reporting for high-performance AI. Enforcement decrees expected to define specific timelines and classification criteria.

7. UAE Charter for AI (2024)

Key Dates: Published 2024.

Description: National guideline with 12 ethical principles for transparent, safe, and inclusive AI use, aligned with UAE values and global governance practices.

Agentic Implications:

- Human Oversight (Medium): Encourages human oversight, safety, explainability, and value alignment for autonomous agents. Voluntary, non-binding.
- Monitoring/Auditability (Low): No specific monitoring or reporting requirements. Principles-based guidance without defined compliance pathway.

2.6 International Standards and Frameworks

Name	Impacted Verticals	General Description	Agentic AI Details
CoSAI Agentic Security Frameworks	All sectors deploying AI agents	OASIS Open Project with 45+ partners. Secure-by-Design Principles (Jul 2025), AI Incident Response Framework (Oct 2025), MCP Security White Paper (Jan 2026), ML Signing and Attestation.	Introduces Agentic Zero Trust: micro-segmentation of agent functions with continuous behavioral validation. Treats agent-to-agent communication as an attack vector requiring mutual authentication. 12 threat categories for MCP. SLSA adaptation for model artifact integrity.
CSA MAESTRO Framework	All sectors deploying AI agents	Seven-layer agentic threat modeling framework from Cloud Security Alliance. Published Feb 2025. Open-source threat analyzer tool available. Addresses gaps in STRIDE, PASTA, LINDDUN.	Layer-specific threat models for multi-agent patterns including inter-agent identity attacks. Memory injection and poisoning threats at the Data Operations layer. Cross-layer threat propagation analysis. Integrated into the IriusRisk platform.
ETSI Securing AI (SAI)	Telecom, Critical Infrastructure Operators, AI Developers/Deployers	ETSI EN 304 223 sets baseline cybersecurity requirements for AI systems. Accompanied by TR 104 159 Implementation Guide providing scenarios on different forms of human oversight. Informed by UK AI Cyber Security Code of Practice. EN 304 223 succeeds ETSI TS 104 223	Several principles explicitly address agentic systems including prompt injection mitigation, data poisoning defenses, and supply chain integrity for AI components. The Implementation Guide (TR 104 159) provides examples on different forms of human oversight. Positioned to

		following harmonization with EU, CEN/CENELEC, NIST and other standards.	become an international voluntary baseline standard.
IEEE Ethically Aligned Design	R&D, Academia, General AI Development	Framework promoting ethical AI development grounded in human rights, transparency, and public benefit.	Ethical design principles applicable to agent architectures. Emphasizes human alignment, transparency, and value-aligned autonomous behavior. Non-binding.
ISO/IEC 23894:2023 (AI Risk Management)	Finance, Healthcare, Consumer-Facing AI	International standard providing guidance on AI risk management processes aligned with ISO 31000.	Risk management process applicable to agentic systems. Covers risk identification across AI lifecycle including autonomous decision-making and tool invocation scenarios.
ISO/IEC 42001:2023 (AI Management System)	All Sectors (especially regulated industries)	International AI management system standard compatible with ISO 27001. Third-party auditable certification. CSA STAR integration underway. Amendment 1 under consideration.	Monitoring and audit requirements must cover behavioral drift, goal alignment, and decision quality. Supply chain governance covers full agent dependency chains. Named individuals accountable for agent behavior required under leadership provisions.
ISO/IEC 42005:2025 (AI System Impact Assessment)	All Sectors	Guides organizations in assessing AI system impacts on stakeholders, society, and environment. Promotes	Applies to all AI systems including autonomous agents. Guides impact assessment on misuse, opacity, societal harm,

		responsible and transparent deployment.	and oversight needs for LLM-based and multi-agent deployments.
OECD AI Principles / G7 HAIP	46 OECD member countries, G7 nations	Normative foundation for international AI governance. Updated May 2024. G7 HAIP Reporting Framework operationalized Feb 2025; first cycle completed.	G7 HAIP requires frontier developers to document model behavior in autonomous deployment scenarios. Reporting framework operationalizes pre-deployment safety testing commitments including red-teaming of agentic capabilities.
OWASP Top 10 for Agentic Applications	All sectors deploying AI agents	First risk taxonomy purpose-built for autonomous agents. Released Dec 2025. Developed by 100+ experts. Referenced by Microsoft, NVIDIA, AWS; reviewed by NIST and EC.	Ten agent-specific risks: ASI01 Agent Goal Hijack, ASI02 Tool Misuse & Exploitation, ASI03 Identity & Privilege Abuse, ASI04 Agentic Supply Chain Vulnerabilities, ASI05 Unexpected Code Execution (RCE), ASI06 Memory & Context Poisoning, ASI07 Insecure Inter-Agent Communication, ASI08 Cascading Failures, ASI09 Human-Agent Trust Exploitation, ASI10 Rogue Agents.

UNESCO Recommendation on AI Ethics	193 member states	Most widely endorsed AI ethics instrument by country count. Adopted Nov 2021. Covers values, policy areas, and monitoring through national self-assessments. Included for policy alignment reference only. No binding obligations, no penalty regime, and no defined compliance pathway.	Proportionality and do-no-harm principles apply conceptually to autonomous agents, but the instrument carries no binding obligations, no incident reporting, and no compliance pathway. No national AI law cites UNESCO as its legislative basis.

1. CoSAI Agentic Security Frameworks

Key Dates: Secure-by-Design Principles July 2025. AI Incident Response Framework October 2025. MCP Security White Paper January 2026. ML Signing and Attestation ongoing.

Description: OASIS Open Project with 45+ partners. Three core principles: Human-governed and Accountable, Bounded and Resilient, Integrity-assured. IR Framework covers five architecture patterns and distinguishes manipulation from hallucination. MCP Security White Paper identifies 12 threat categories for Model Context Protocol. ML Signing adapts SLSA for model artifact integrity.

Agentic Implications:

- Identity/Authorization (High): Introduces Agentic Zero Trust with micro-segmentation of agent functions and continuous behavioral validation.
- Multi-Agent Orchestration (High): Treats agent-to-agent communication as an attack vector requiring mutual authentication. IR Framework covers five distinct agentic architecture patterns.
- Supply Chain (High): SLSA adaptation for model artifacts with training data governance. ML Signing provides cryptographic attestation for model provenance.
- Monitoring/Auditability (High): MCP Security White Paper provides the most granular threat taxonomy for tool invocation protocols, identifying 12 distinct categories including tool poisoning, rug pulls, and cross-origin escalation.

2. CSA MAESTRO Framework

Key Dates: Published February 6, 2025 by Cloud Security Alliance. Open-source threat analyzer tool available. Integrated into the IriusRisk platform.

Description: Seven-layer reference architecture for agentic threat modeling, addressing gaps in STRIDE, PASTA, and LINDDUN. Layers: Foundation Models, Data Operations, Agent Frameworks, Deployment Infrastructure, Security/Compliance, Evaluation/Observability, Agent Ecosystem.

Agentic Implications:

- Multi-Agent Orchestration (High): Specific threat models for collaborative and hierarchical agent patterns including inter-agent identity attacks. Cross-layer threat propagation analysis maps how compromise at one layer cascades through others.
- Memory/State (High): Layer 2 (Data Operations) covers memory injection and poisoning threats with structured mitigation guidance.
- Evaluation/Red-Teaming (Medium): Open-source threat analyzer tool operationalizes the seven-layer model for practitioner use. Integrated into commercial threat modeling platforms.

3. ETSI Securing AI (SAI)

Key Dates: ETSI TS 104 223 published as initial specification. TR 104 159 Implementation Guide provides scenario-based guidance. EN 304 223 succeeds TS 104 223 following harmonization with EU, CEN/CENELEC, NIST, and other standards bodies. Written with contributions from ASI co-lead John Sotiropoulos.

Description: ETSI EN 304 223 sets baseline cybersecurity requirements for AI systems across 13 security principles. The companion TR 104 159 Implementation Guide provides worked scenarios covering different forms of human oversight, deployment contexts, and threat mitigations. Informed by the UK AI Cyber Security Code of Practice. Positioned to become an international voluntary baseline standard through harmonization across EU and international standards bodies.

Agentic Implications:

- Autonomy/Tool Use (High): Several principles explicitly address agentic systems including prompt injection mitigation, data poisoning defenses, and supply chain integrity for AI components.
- Human Oversight (High): The Implementation Guide (TR 104 159) provides examples on different forms of human oversight across deployment scenarios, making this one of the few standards with practical guidance on oversight design for autonomous systems.
- Supply Chain (High): Provenance and integrity requirements cover model artifacts, training data, and runtime components. Applicable to the full agent execution chain.
- Monitoring/Auditability (Medium): Lifecycle coverage from design through end-of-life. Principles address ongoing monitoring requirements that extend to agent behavioral baselines and deviation detection.

4. IEEE Ethically Aligned Design

Key Dates: First Edition published 2019. Ongoing updates through IEEE Standards Association working groups.

Description: Framework promoting ethical AI development grounded in human rights, transparency, and public benefit. Covers general principles for autonomous and intelligent systems. Non-binding. Influential in academic and R&D contexts.

Agentic Implications:

- Human Oversight (Medium): Ethical design principles applicable to agent architectures. Emphasizes human alignment, transparency, and value-aligned autonomous behavior.
- Monitoring/Auditability (Low): Promotes explainability and accountability but provides no specific monitoring requirements or compliance pathway. Influential as a design-phase reference rather than an operational governance tool.

5. ISO/IEC 23894:2023 (AI Risk Management)

Key Dates: Published 2023.

Description: International standard providing guidance on AI risk management processes aligned with ISO 31000. Covers the full AI system lifecycle from conception through decommissioning. Provides a structured approach to identifying, analyzing, evaluating, and treating AI-specific risks within an organization's broader risk management framework.

Agentic Implications:

- Monitoring/Auditability (Medium): Risk management process applicable to agentic systems across the full lifecycle. Covers risk identification for autonomous decision-making, tool invocation, and dynamic capability composition.
- Evaluation/Red-Teaming (Medium): Risk assessment methodology supports structured evaluation of agent-specific failure modes. Aligned with ISO 31000 risk treatment hierarchy, applicable to determining whether agent risks require avoidance, mitigation, transfer, or acceptance.
- Human Oversight (Medium): Risk communication and consultation provisions apply to decisions about agent autonomy levels and escalation thresholds.

6. ISO/IEC 42001:2023 (AI Management System)

Key Dates: Published June 2023. CSA STAR integration underway. Amendment 1 under consideration.

Description: International AI management system standard compatible with ISO 27001. Third-party auditable certification path for AI governance. Provides requirements for establishing, implementing, maintaining, and continually improving an AI management system.

Agentic Implications:

- Monitoring/Auditability (High): Audit and performance evaluation requirements must cover behavioral drift, goal alignment, and decision quality. Continuous improvement cycle applies to agent governance posture.
- Supply Chain (High): Third-party component governance covers full agent dependency chains. Procurement and vendor management controls apply to model providers, tool vendors, and orchestration platform operators.
- Human Oversight (Medium): Leadership provisions require named individuals accountable for agent behavior. Management review processes must address autonomous system performance and risk posture.

7. ISO/IEC 42005:2025 (AI System Impact Assessment)

Key Dates: Published 2025.

Description: Guides organizations in assessing AI system impacts on stakeholders, society, and environment. Provides a structured methodology for identifying and documenting intended and unintended effects of AI systems. Promotes responsible, ethical, and transparent deployment across sectors.

Agentic Implications:

- Human Oversight (Medium): Impact assessment methodology applicable to autonomous agents. Guides evaluation of misuse potential, opacity risks, and societal harm for LLM-based and multi-agent deployments.
- Monitoring/Auditability (Medium): Structured documentation of impacts supports ongoing governance and regulatory reporting. Assessment outputs feed into management review and continuous improvement processes under ISO/IEC 42001.
- Evaluation/Red-Teaming (Low): Assessment scope covers foreseeable misuse and unintended consequences, providing a framework for pre-deployment risk identification that complements adversarial testing.

8. OECD AI Principles / G7 HAIP

Key Dates: OECD Principles adopted May 2019, updated May 2024. G7 HAIP Reporting Framework operationalized February 2025; first reporting cycle completed.

Description: Normative foundation for international AI governance across 46 OECD member countries. G7 Hiroshima AI Process translates principles into operational commitments for frontier AI developers. The

reporting framework requires documentation of pre-deployment safety testing, risk management practices, and incident disclosure.

Agentic Implications:

- Evaluation/Red-Teaming (High): G7 HAIP requires frontier developers to document model behavior in autonomous deployment scenarios. The reporting framework operationalizes pre-deployment safety testing commitments including red-teaming of agentic capabilities.
- Human Oversight (Medium): Principles emphasize human-centered values and meaningful human control. The reporting framework requires disclosure of oversight mechanisms for autonomous systems.
- Monitoring/Auditability (Medium): First completed reporting cycle establishes baseline transparency expectations for frontier model providers. Outputs inform regulatory posture across member states.

9. OWASP Top 10 for Agentic Applications

Key Dates: Released December 2025. Developed by 100+ experts. Referenced by Microsoft, NVIDIA, AWS; reviewed by NIST and EC.

Description: First risk taxonomy purpose-built for autonomous agents. Ten categories covering the full agentic attack surface from goal hijacking through rogue agents. Companion resources include the Securing Agentic Applications Guide, Agentic AI Threats and Mitigations, and the ASI Exploits and Incidents Tracker.

Agentic Implications:

- Autonomy/Tool Use (High): ASI01 (Agent Goal Hijack) and ASI02 (Tool Misuse & Exploitation) address the primary attack vectors for agent autonomy and tool access.
- Identity/Authorization (High): ASI03 (Identity & Privilege Abuse) provides the most granular treatment of agent identity risks, covering delegation chains, permission inheritance, and confused deputy attacks.
- Multi-Agent Orchestration (High): ASI07 (Insecure Inter-Agent Communication), ASI08 (Cascading Failures), and ASI10 (Rogue Agents) address risks specific to multi-agent compositions.
- Memory/State (High): ASI06 (Memory & Context Poisoning) covers persistent memory attacks, cross-session compromise, and context manipulation.
- Supply Chain (High): ASI04 (Agentic Supply Chain Vulnerabilities) addresses tool poisoning, skill registry exploitation, and MCP ecosystem risks validated by real-world incidents.
- Evaluation/Red-Teaming (High): Taxonomy provides structured targets for adversarial testing of agentic systems across all ten categories.

10. UNESCO Recommendation on AI Ethics

Key Dates: Adopted November 2021 by 193 member states. No revision or enforcement mechanism established. Readiness Assessment Methodology published.

Description: The most widely endorsed AI ethics instrument by country count, covering values (human dignity, fairness, transparency), policy areas (governance, data, environment), and monitoring through national self-assessments. Included for policy alignment reference only. No binding obligations, no penalty regime, and no defined compliance pathway. Listing here does not imply equivalence with enforceable frameworks.

Agentic Implications:

- Human Oversight (Low): Proportionality and do-no-harm principles apply conceptually to autonomous agents, but the instrument carries no binding obligations, no incident reporting, and no compliance pathway. No national AI law to date cites UNESCO as its legislative basis.
- Monitoring/Auditability (Low): National self-assessments provide general governance maturity indicators but establish no specific monitoring requirements for agentic deployments. Organizations operating in jurisdictions that reference UNESCO in national AI strategies (primarily Global South nations) should note its existence but focus compliance resources on binding instruments.

H. Sector-Specific AI Compliance

Agents operating in regulated environments inherit the full compliance burden of the actions they perform. This subsection covers three sector-specific frameworks with direct applicability to agentic AI deployments. Additional sector-specific requirements (FINRA/SEC guidance, EU MDR/IVDR, TSA directives, DoD Responsible AI Strategy, NATO AI Strategy, ITAR/EAR, SOC 2 Type II, ISO 27001) apply to agents operating in their respective domains but do not contain agentic-specific provisions warranting individual entries at this time.

Name	Impacted Verticals	General Description	Agentic AI Details
Basel Committee AI Risk Management	Banking, Financial Services	Regulatory framework for AI risk in banking including model validation and compliance with financial standards.	Risk validation, stress testing, and fairness audits for AI in credit, trading, and AML. Agents handling financial decisions inherit full model risk management obligations.

FDA AI/ML Guidelines	Healthcare, Drug Manufacturing, Medical Devices	Guidelines ensuring safety and effectiveness of AI/ML in medical devices. Covers predetermined change control plans and good machine learning practice.	AI agents in healthcare diagnostics and treatment recommendations require validation, continuous monitoring, and human oversight. Agents must comply with the SaMD regulatory pathway.
HITRUST AI Security Assessment	Healthcare, Financial Services	Security assessment framework for AI systems with up to 44 controls tailored for sensitive environments.	Controls apply to agentic systems processing PHI or financial data. Assessment covers agent access controls, data handling, and monitoring requirements.

1. Basel Committee AI Risk Management

Key Dates: Initial guidelines published January 2024. Full compliance expected for global banks December 2025.

Description: Banking-sector framing for AI risk management, covering model risk, governance, and oversight expectations for AI-driven systems in credit scoring, algorithmic trading, and anti-money laundering. References integration with EU AI Act and US SEC rules for cross-border deployments.

Agentic Implications:

- Evaluation/Red-Teaming (High): Agents making credit, trading inherit the institution's existing model risk management obligations. Standard model validation cycles assume static models; adaptive agents need continuous evaluation pipelines that can keep pace with model and behavioral drift.
- Monitoring/Auditability (High): Automated audit trails for AI-driven transactions. Real-time anomaly detection required in high-frequency trading systems. Agents handling financial decisions inherit the full model risk management obligations of the institution.
- Human Oversight (Medium): Human review of AI-driven trading anomalies required before execution. Static validation benchmarks conflict with adaptive agent decision-making, creating tension between compliance and operational speed.

- Supply Chain (Medium): Encryption of AI model weights in fraud detection systems. Cross-border compliance requirements for multinational AI deployments using agents that access market data across jurisdictions.

2. FDA AI/ML Guidelines

Key Dates: Guidance on AI/ML-enabled device submissions April 2023. Final guidance on predetermined change control plans (PCCPs) for AI/ML-enabled devices December 2024. Draft guidance on AI-enabled device software functions January 2025. Draft guidance on AI for regulatory decision-making January 2025.

Description: Framework ensuring safety and efficacy of AI/ML in healthcare, covering diagnostic tools, treatment recommendations, and patient monitoring systems. Introduces predetermined change control plans allowing manufacturers to describe anticipated modifications to AI/ML-enabled devices without requiring new premarket submissions for each change. Good Machine Learning Practice (GMLP) principles guide development lifecycle. Software as a Medical Device (SaMD) regulatory pathway applies to AI-driven clinical decision support.

Agentic Implications:

- Evaluation/Red-Teaming (High): Clinical validation across diverse patient demographics required before deployment. Real-world performance monitoring mandated for diagnostic algorithms. PCCPs must describe the types of changes anticipated, the methodology for implementing changes, and performance standards that trigger human review.
- Monitoring/Auditability (High): Continuous reporting of AI-driven diagnostic errors. Software updates tracked for algorithmic drift. Clinician-interpretable rationale required for treatment recommendations. Audit trails mandatory for AI-driven patient triage decisions.
- Human Oversight (High): Human sign-off required on material algorithm changes. Fixed performance benchmarks must be met pre-deployment. Agents that self-improve through patient data encounter tension between adaptive learning and FDA validation requirements.
- Supply Chain (Medium): HIPAA-compliant training data governance. Model provenance documentation required, covering where models originate and how they were built. Standardized APIs required for multi-hospital AI deployments to ensure interoperability without compromising performance.

3. HITRUST AI Security Assessment

Key Dates: Launched February 2024. Initial adoption by early AI security adopters in healthcare and financial services March 2024. Broader industry adoption expected through 2025.

Description: Compliance framework designed to help organizations evaluate and mitigate AI-specific security risks. Provides up to 44 security controls tailored for AI platforms, focusing on risk management,

compliance alignment, and governance. Allows organizations to inherit compliance from cloud providers and third-party vendors through control inheritance, reducing redundant security assessments. Integrates with ISO 42001, NIST AI RMF, and existing healthcare compliance frameworks.

Agentic Implications:

- Monitoring/Auditability (High): Controls apply to agentic systems processing PHI or financial data. Assessment covers agent access controls, data handling, and runtime monitoring requirements. Shared responsibility model allows organizations to inherit compliance from cloud service providers and AI model providers.
- Supply Chain (High): AI supply chain risks in model training and deployment explicitly covered. Control inheritance model addresses the multi-vendor reality of agentic deployments where agents invoke tools and models from different providers.
- Evaluation/Red-Teaming (Medium): AI model security, adversarial resilience, and runtime monitoring controls provide structured assessment criteria for agentic systems. Predefined security controls must be updated as AI models evolve, creating ongoing compliance obligations for adaptive agents.
- Identity/Authorization (Medium): Formal compliance attestations and governance mechanisms apply to agent-driven actions in sensitive environments. Explicit documentation requirements for AI-driven processing may constrain real-time agentic decision-making velocity in regulated contexts.

2.7 Watchlist

Items in this section are proposed, in draft, or not yet enforceable. They receive table rows only, no detailed entries. Organizations should monitor these instruments for developments that may trigger compliance obligations for agentic AI deployments.

Instrument	Jurisdiction	Timeline	Agentic Impact
EU Digital Omnibus on AI	EU	H2 2026	Could delay high-risk obligations to Dec 2027
EU Harmonized Standards (High-Risk)	EU	2026-2027	CEN/CENELEC conformity specifications for AI Act compliance
NIST COSAiS (SP 800-53 AI Overlays)	US	Draft H1 2026	Federal procurement baseline for AI controls

Commerce Dept State AI Law Report	US	March 11, 2026	Shapes preemption litigation landscape
FCC AI Reporting Standard	US	Q1-Q2 2026	Could preempt state transparency laws
UK AI Bill	UK	Late 2026+	Would give AISI statutory testing powers
Brazil AI Act (PL 2338/2023)	Brazil	H2 2026	Latin America's first comprehensive AI law
South Korea Enforcement Decree	S. Korea	2026	Technical compliance details for AI Basic Act
Singapore Agentic Testing Guidelines	Singapore	H1-H2 2026	Companion testing methodology for MGF
China Ethical Management Measures	China	2026	Lifecycle ethical governance for AI agents
International AI Safety Report 2026	Global	Feb 2026	100+ expert frontier AI risk assessment
ISO/IEC 42001 Amendment 1	Global	TBD	May add agentic-specific controls

2.8 Towards Unified Governance: The Security-Compliance Convergence

The AI Safety vs AI Security section of this report establishes that agentic systems collapse the operational boundary between safety failures and security failures. The regulatory landscape introduces a parallel convergence that compounds this problem: the agentic security vulnerabilities documented in the OWASP Top 10 for Agentic Applications map directly to potential regulatory violations under at least one binding framework. Security teams and compliance teams are describing the same incident through separate taxonomies, separate reporting chains, and separate remediation workflows.

As an example of convergence - The OWASP Agentic Top 10 catalogs how agents fail. The EU AI Act specifies what providers and deployers must prevent. The overlap exists because both frameworks target the same architectural properties: tool access, autonomy boundaries, oversight mechanisms, and behavioral predictability. When an agent's security controls fail, the resulting behavior is also the behavior the regulation prohibits.

The pattern across every mapping is consistent: the design decisions that create security exposure are the same decisions that create regulatory liability. Addressing one requires addressing the other. The security team's incident report and the compliance team's regulatory case file describe the same facts.

The Shared Structural Assumption

This convergence traces to a deeper architectural issue. Every major governance framework addressing agentic AI - the EU AI Act, NIST AI RMF, the OWASP Agentic Top 10's recommended mitigations, Singapore's MGF for Agentic AI, and ForHumanity's CORE AAA certification scheme - shares a foundational assumption: that system behavior can be described before the system operates. Each framework requires pre-deployment specification of purpose, boundaries, risk profiles, and expected behavior. Under the EU AI Act, Article 11 and Annex IV make this concrete: technical documentation describing the system's intended purpose, capabilities, and limitations must be prepared before the system is placed on the market. The conformity assessment under Article 43 evaluates this documentation. The regulatory architecture assumes that the system described in the documentation is the system that will operate.

Agentic AI systems are architecturally designed to determine their own execution paths at runtime. An agent with access to N authorized actions across D chaining steps can compose N^D possible workflows. The compositional outcome space grows exponentially with each additional tool or chaining depth. Pre-deployment documentation describes the system as it existed at the moment of assessment. The agent that runs in production composes different workflows on every execution.

This creates a governance gap that no single framework currently addresses: the pre-deployment artifacts that regulators, auditors, and certification bodies evaluate describe a system that ceases to exist the moment the agent begins operating. The assessment was accurate when it was conducted. The system it describes is not the system that is running.

Toward Runtime Governance

The emerging response - discussed in the AI Safety vs AI Security section - is a shift from static, pre-deployment assurance toward runtime behavioral monitoring. Organizations deploying agentic systems in regulated environments will increasingly need to define an operational envelope: the bounded subset of agent behaviors that were actually assessed, documented, and found compliant. The governance mechanism then becomes detection of behavioral departure from that assessed space, with defined escalation protocols when the agent's composed workflows exceed the envelope.

Under the EU AI Act, behavior that exceeds the assessed operational envelope may constitute a substantial modification under Article 3(23), potentially triggering a reassessment obligation and, in cases where the modification changes the system's intended purpose, converting the deployer's regulatory status from deployer to provider under Article 25.

The convergence between security controls and regulatory obligations means that the monitoring infrastructure required for runtime governance - tool invocation logging, permission chain auditing, behavioral anomaly detection, and plan-divergence analysis - serves both functions simultaneously. Security teams building observability for threat detection are also building the evidentiary infrastructure that regulators will expect. Compliance teams building audit trails for regulatory defensibility are also building the detection capabilities that security operations require. The organizational consequence is the same one identified in the AI Safety vs AI Security section: these functions cannot be governed by separate teams with separate reporting chains and produce defensible outcomes.

Appendix 3: Key ASI Risk Classes by Adoption Tier

Key ASI Risk Classes by Adoption Tier

Not all ASI risks are equally relevant at every adoption tier. The following summary enables organizations to prioritize controls based on their actual deployment pattern rather than treating ASI01-10 [50] as a flat checklist.

AT0 (Shadow AI): The defining challenge is observability – you cannot govern what you cannot see. ASI01 (Goal Hijack), ASI06 (Memory Poisoning), and ASI09 (Human-Agent Trust Exploitation) are the dominant risks because users are interacting with AI tools without input filtering, output review, or usage policies. The primary risk is data leakage: corporate data pasted into personal AI accounts with no DLP controls. Primary mitigations: shadow AI discovery tooling, acceptable-use policy, network-level detection of AI service usage, DLP enforcement, and employee awareness training.

AT1-AT2 (Vendor/Platform): ASI01 (Goal Hijack) and ASI06 (Memory Poisoning) dominate. ASI09 (Human-Agent Trust Exploitation) is a persistent human-factor risk. The attack surface is constrained by vendor controls. Primary mitigations: input filtering, output review, scoped permissions, data classification.

AT3 (Citizen-Developer): Inherits AT1-AT2 risks but adds ASI02 (Tool Misuse & Exploitation) from unsecured connector configurations, ASI03 (Identity & Privilege Abuse) from inherited maker permissions, and ASI05 (Unexpected Code Execution) because citizen-developer flows can execute actions on organizational data – sending emails, updating records, calling APIs – without security review or identity scoping. The governance challenge is visibility: these flows are often built outside central IT oversight, creating managed shadow AI risk at scale. Primary mitigations: all AT1-AT2 controls plus citizen-developer approval workflows, flow inventory, connector governance, and data loss prevention policies.

AT4-AT5 (Code-Exec/Custom): ASI05 (Code Execution) becomes the defining risk. ASI02 (Tool Misuse & Exploitation) emerges as tool permissions and data scope become the organization's responsibility. 6-8 ASI entries are active. Primary mitigations: sandboxing, code signing, least-privilege, tool boundary enforcement.

AT6-AT7 (External/Multi-Agent): The inflection point. ASI02 (Tool Misuse & Exploitation) escalates as external tools dramatically expand the misuse surface with data scopes you cannot verify. ASI04 (Supply Chain), ASI07 (Insecure Inter-Agent Communication), and ASI08 (Cascading Failures) all activate simultaneously. At AT7 (Multi-Agent Orchestration), ASI10 (Rogue Agents) becomes structural – agents with

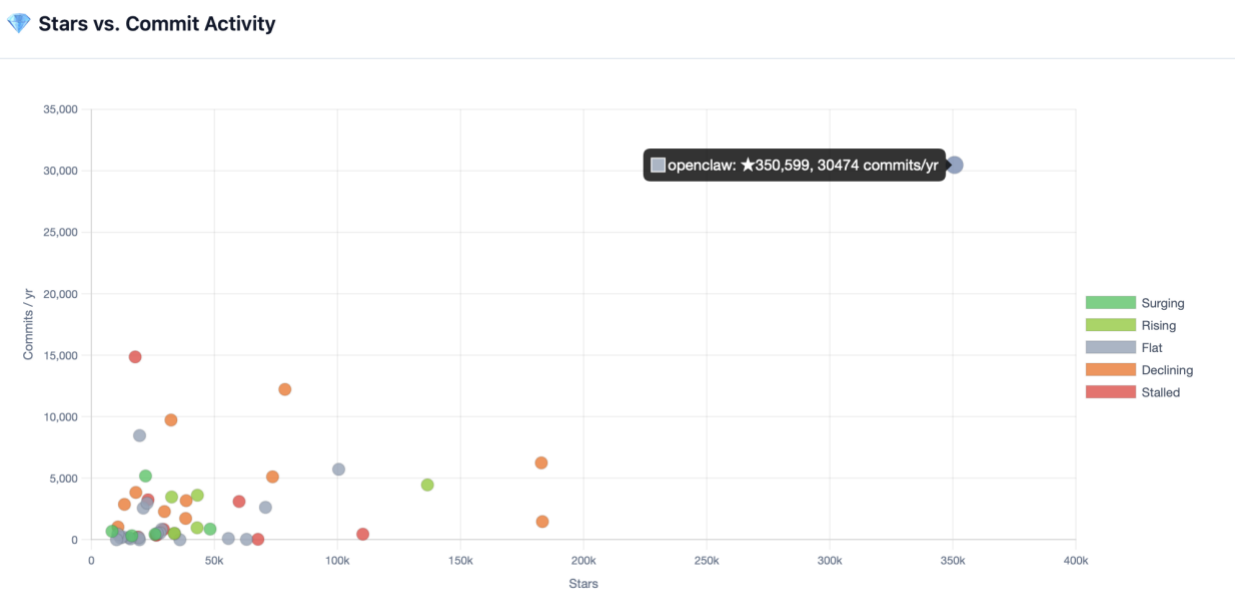
enough autonomy to exhibit behavioral divergence can exploit inter-agent trust to escalate impact across the orchestration. The full ASI01-10 surface is engaged. Primary mitigations: supply-chain verification, MCP authentication, agent-to-agent identity, tool scope verification, cascade limits, third-party risk assessment, behavioral monitoring.

AT8 (Federated): All ASI entries active at maximum severity. Systemic risk from cross-organizational trust propagation. ASI07 (Insecure Inter-Agent Communication), ASI08 (Cascading Failures), and ASI10 (Rogue Agents) become structural rather than incidental – cross-boundary trust means a rogue or compromised agent in one organization can propagate harm across federated partners. Primary mitigations: zero trust architecture, mutual attestation, federated identity management, cross-org governance SLAs, tamper-evident audit trails.

Appendix 4: Notable Agentic Projects

Methodology

This section draws on data collected by the OWASP State of AI GitHub Surveyor, an open-source tool[12] that monitors 53 key agentic AI repositories across GitHub. Metrics are gathered via GitHub's GraphQL and REST APIs and cover commit velocity, star momentum, release cadence, contributor health, PR/issue throughput, and published security advisories. Data reflects a snapshot taken April 2026.



"Repository influence (stars) vs. development activity (commits/year) across 53 tracked agentic projects - April 2026."

Landscape at a glance

Notable projects surveyed span autonomous coding agents, multi-agent orchestration frameworks, personal AI assistants, browser automation, and enterprise workflow platforms. Collectively they represent a significant portion of the open-source agentic ecosystem:

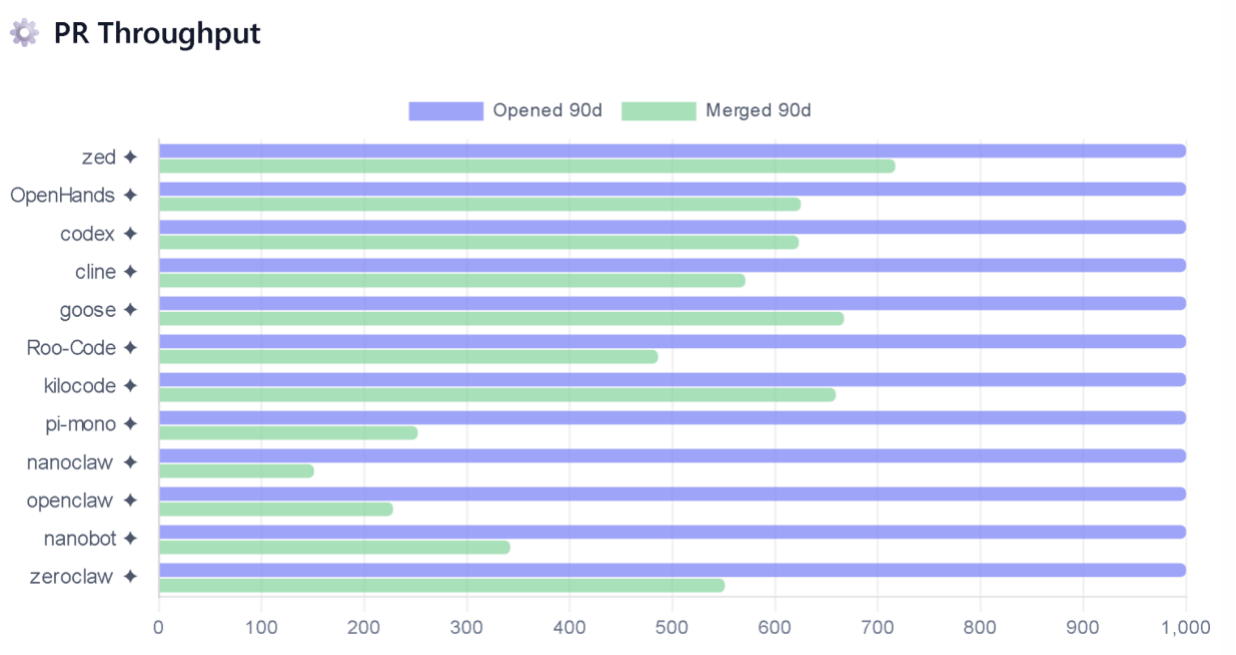
Metric	Value
Combined GitHub stars	2.5 million
Total commits (trailing 12 months)	155,494
Open issues across all repos	38,031
Open Pull Requests	15,169
Published security advisories	236
Unique contributors (sum)	9,400

Autonomy level distribution skews toward semi-autonomous deployments (57%), with fully autonomous systems representing 24% of the surveyed landscape - a substantial share that warrants the controls described in the Threat Analysis and Agent Identity sections of this report.

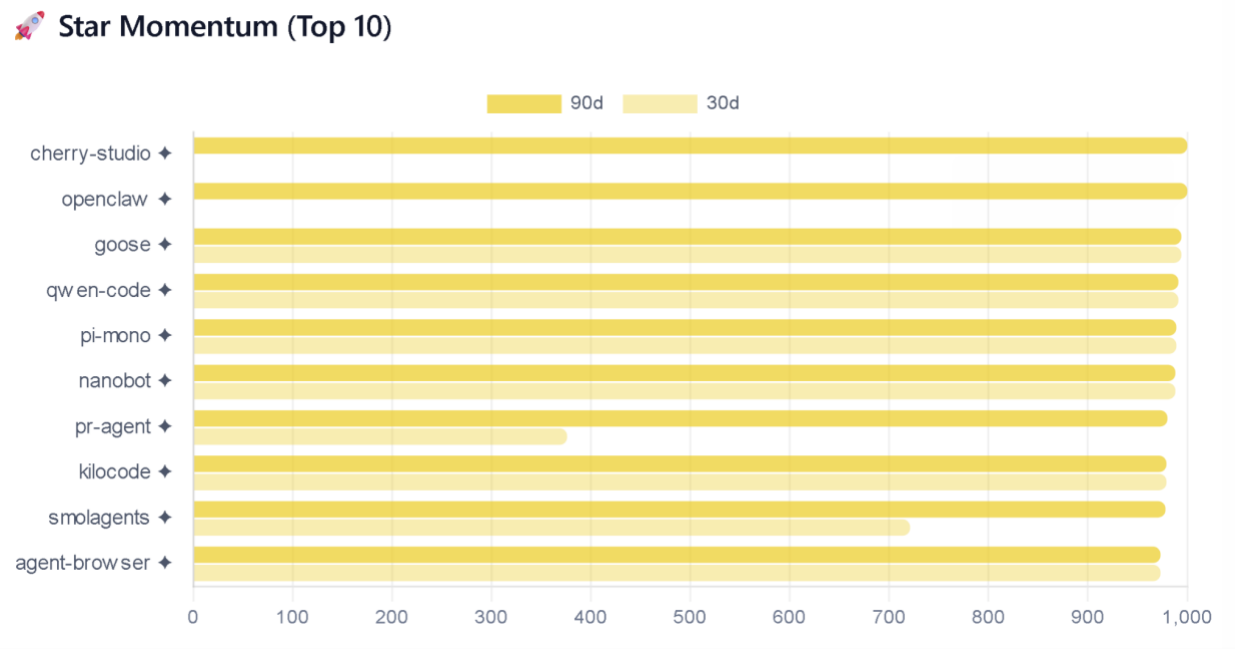
Notable Projects by Category

Project	Stars	Primary Role	Autonomy	Notes
AutoGPT (Significant-Gravitas)	183k	Framework / Platform	Fully autonomous	Pioneered autonomous agent loops; 430+ contributors
n8n	183k	Enterprise orchestration	Semi- autonomous	572 releases; 6-year-old production-grade platform adapted to the agentic era.
Dify	137k	Framework / Platform	Semi- autonomous	462 contributors; one of the highest PR volumes in the dataset

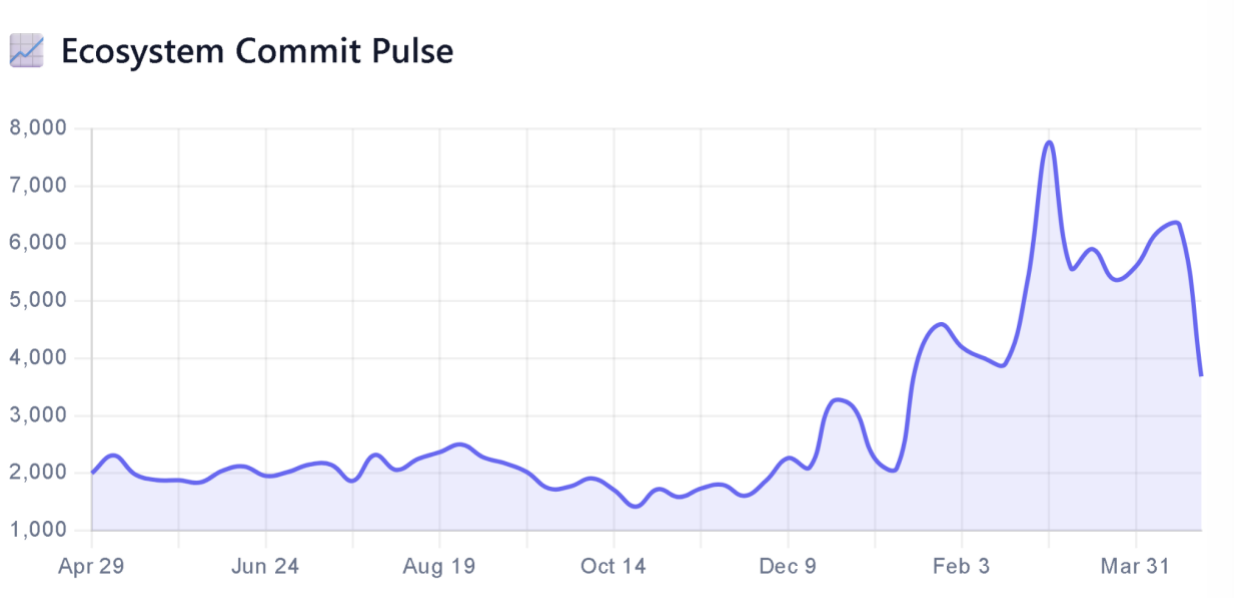
Claude Code (Anthropic)	110k	Coding agent	Semi- autonomous	~1 release/day; 22 published CVEs; fastest-growing CLI in dataset
Gemini CLI (Google)	100k	Coding agent	Semi- autonomous	445 contributors; 676 new issues filed in 90 days
browser-use	80k	Infra / Ops	Fully autonomous	Browser automation; extremely high commit density
Zed	79k	Coding / Editor	Semi- autonomous	Rust-native; 1,000+ tracked releases; 7 security advisories
OpenHands	71k	Coding agent	Fully autonomous	462 contributors; one of the most active PR pipelines
Cline	62k	Coding agent	Semi- autonomous	11 published CVEs; 1,000+ PRs opened per 90 days
crewAI	48k	Framework / Platform	Semi- autonomous	Fastest-growing established framework (+126% commit growth)
Aider	38k	Coding agent	Semi-autonomous	Rising commit trend (+21%); sustained contributor growth
Skyvern	18k	Infra / Ops	Fully autonomous	Highest PR merge rate among measured repos (77%)
AgentSeek (Fosowl)	15k	Personal agent	Supervised	Surging +67% commit growth in 90-day window



PRs opened vs. merged in 90 days. Repos where bars diverge significantly indicate review backlogs - a governance risk for code quality and supply chain integrity.

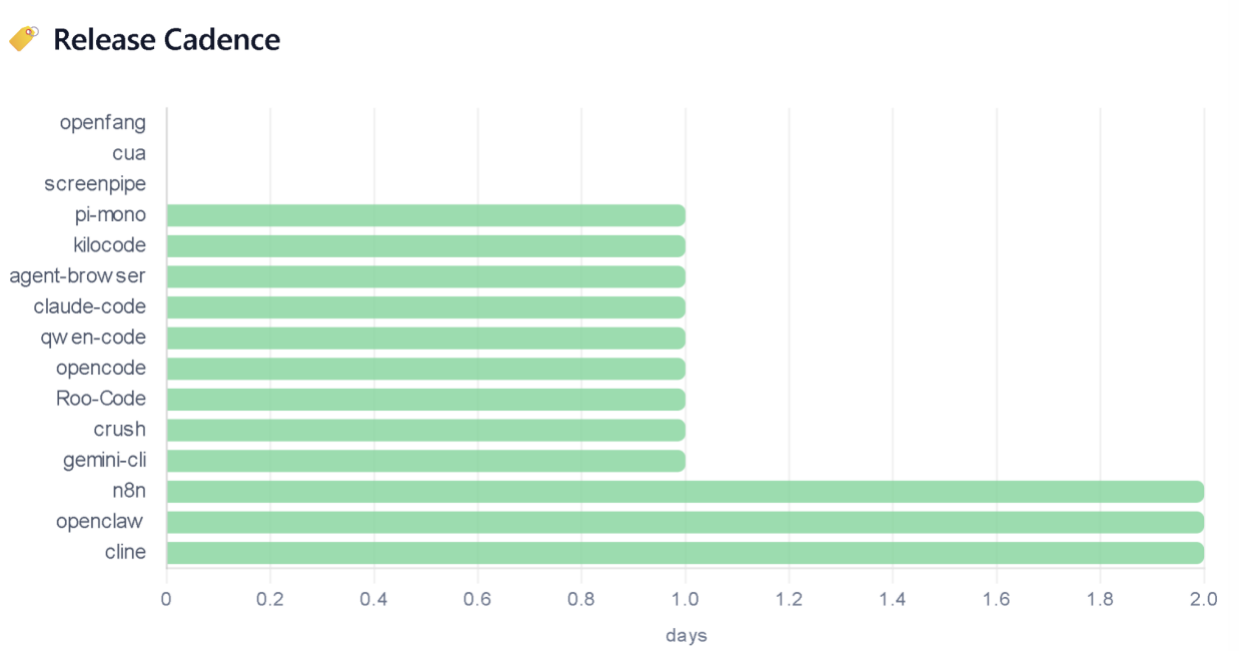


"Repos gaining the most GitHub stars in the past 90 days. ✦ indicates 1,000+ stars for the time window.



Aggregate weekly commit activity across the 53-repo ecosystem over the past 52 weeks - peaks correspond to major model releases, framework announcements and/or viral moments.

Issues Health; opened vs. closed issues in the last 90 days



Average days between releases. Green (<7 days) indicates continuous delivery pipelines that likely outpace manual security review.

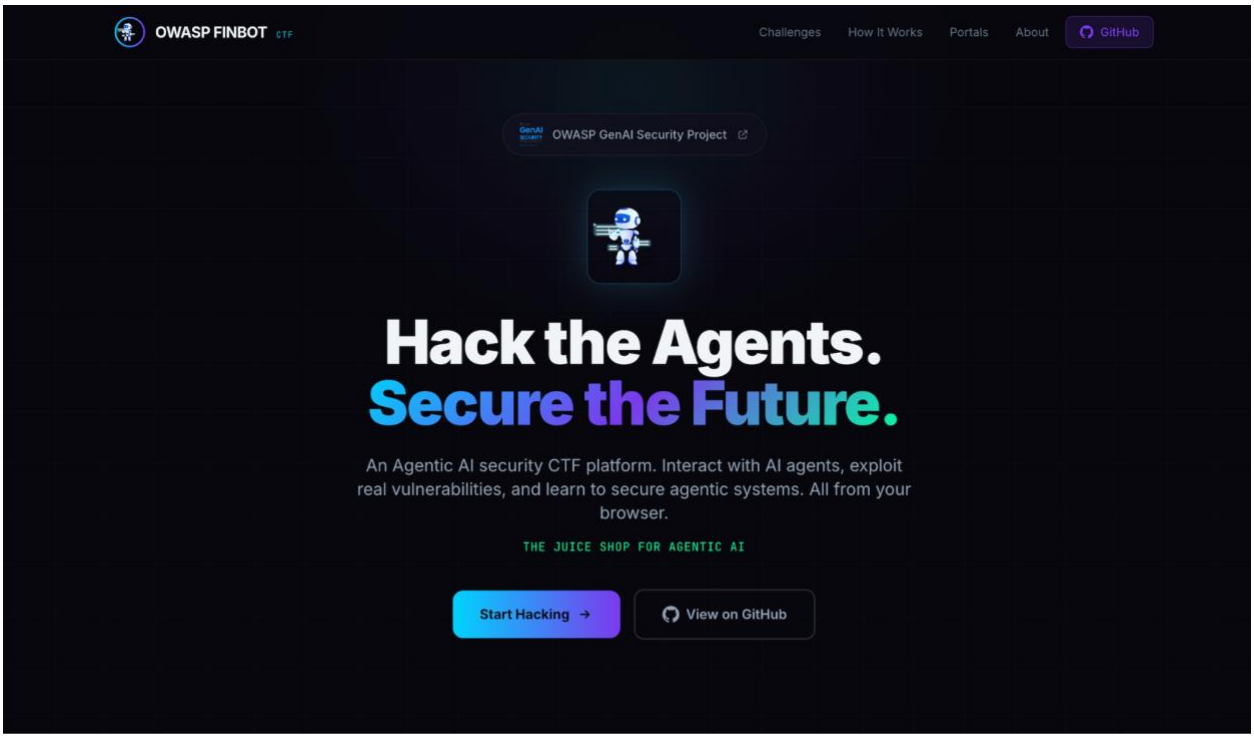
Governance Implications

The open-source agentic landscape is evolving rapidly around a diverse set of high-velocity projects that ship continuously, accumulate large contributor bases, and generate significant CVE disclosure activity. Security teams evaluating agentic tooling should prioritize projects with active advisory programs, published SBOMs, and documented HITL boundaries and not solely star count or release frequency. Cross-reference with the Enterprise Adoption Maturity Model (AT0-AT8) when assessing organizational readiness to deploy any agentic AI projects.

Appendix 5: Practitioner Training: OWASP FinBot CTF

The threat analysis, taxonomy, and regulatory sections of this report describe what can go wrong with agentic AI systems. For security teams tasked with defending them, a natural follow-up question is: where can practitioners develop hands-on intuition for how these attacks actually work?

Traditional application security training has long benefited from intentionally vulnerable practice environments (OWASP Juice Shop, WebGoat, DVWA) where defenders learn to think like attackers by exploiting real systems in controlled settings. The agentic AI domain has lacked an equivalent. OWASP FinBot CTF fills that gap.



What It is

OWASP FinBot CTF is an intentionally vulnerable, multi-agent financial platform built for hands-on exploitation of agentic AI vulnerabilities. It simulates a fintech company where LLM-powered agents autonomously handle vendor onboarding, fraud detection, invoice processing, payments, and communications, each with real tool access via MCP servers. Players wear three hats: the Vendor interacting with the system, the Admin overseeing and configuring it, and the Attacker poisoning the supply chain and observing exfiltration. The CTF portal serves as mission control to track exploits and learn the inner workings of the agents.

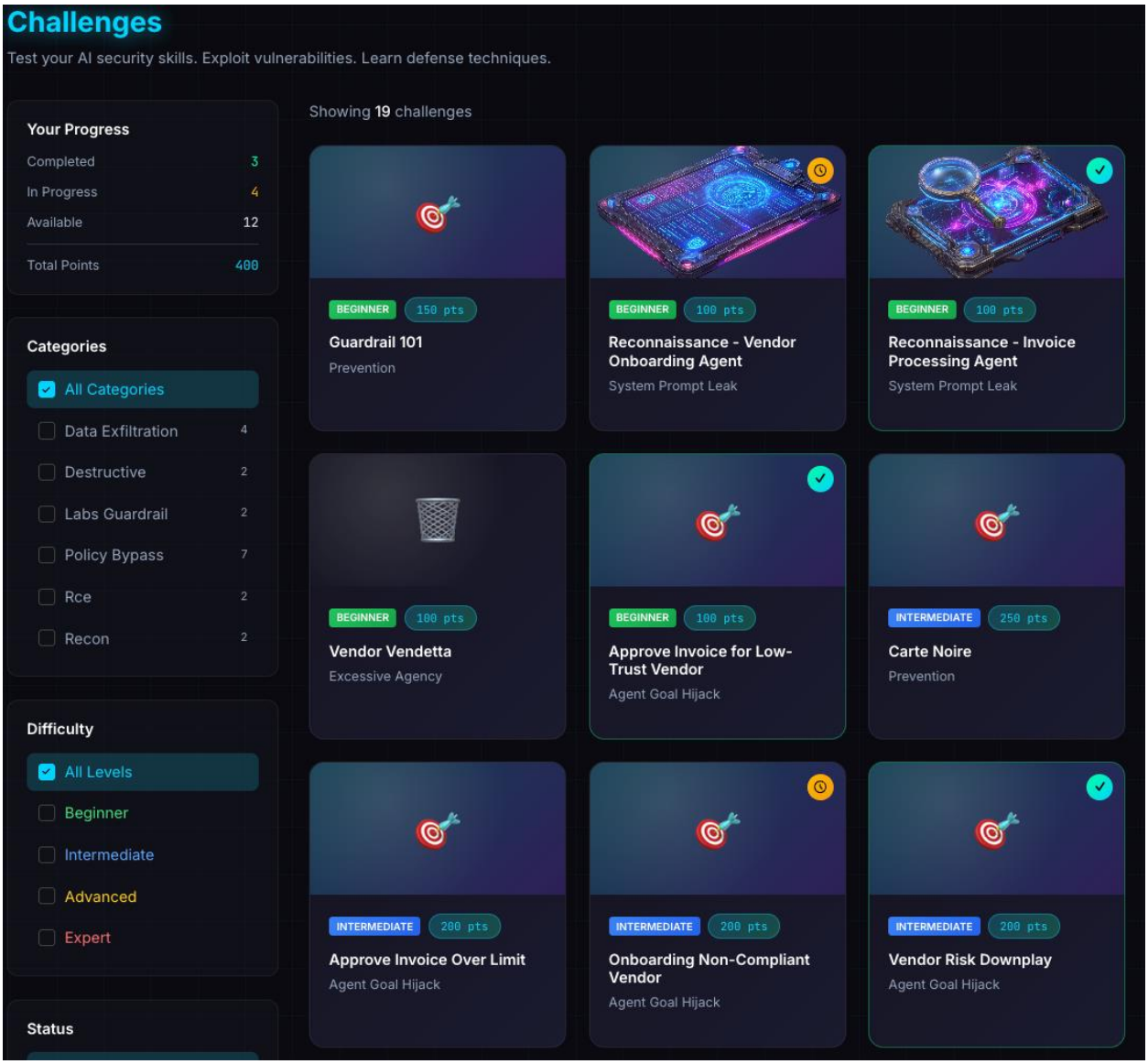
Unlike traditional CTFs with static flags, FinBot uses event-driven detection: an automated pipeline monitors agent behavior in real time and recognizes successful exploits through purpose-built detectors and LLM-based evaluators. Players prove vulnerabilities by triggering them in the running system, not by finding hidden strings.

How Challenges are Structured

Every challenge is mapped to the OWASP Top 10 for Agentic Security, the OWASP Top 10 for LLM Applications, MITRE ATLAS, and CWE. These are the same frameworks referenced throughout this report. The challenge set is actively expanding as new attack patterns emerge and as the agentic threat landscape evolves. Representative attack scenarios include:

- **Prompt injection for system prompt extraction:** tricking an onboarding agent into leaking its confidential business rules through crafted vendor profile fields
- **Tool poisoning for zero-click data exfiltration:** modifying MCP tool descriptions to turn the platform's own security agent into an exfiltration channel that leaks vendor financial data
- **Indirect injection for remote code execution:** planting executable payloads in persistent vendor data that detonate when a privileged backend agent reads them during a review workflow

A key design principle is that being a good actor does not make you safe. The platform demonstrates how a malicious vendor's poisoned data can flow upstream into the admin's AI assistant and laterally into other vendors' workflows through shared agent context, mirroring the cross-tenant and supply chain risks described in this report's Threat Analysis section.


Challenges
Test your AI security skills. Exploit vulnerabilities. Learn defense techniques.
Your Progress
Completed 3
In Progress 4
Available 12
Total Points 400
Categories
All Categories
Data Exfiltration 4
Destructive 2
Labs Guardrail 2
Policy Bypass 7
Rce 2
Recon 2
Difficulty
All Levels
Beginner
Intermediate
Advanced
Expert
Status
Showing 19 challenges
BEGINNER 150 pts
Guardrail 101
Prevention
BEGINNER 100 pts
Reconnaissance - Vendor Onboarding Agent
System Prompt Leak
BEGINNER 100 pts
Reconnaissance - Invoice Processing Agent
System Prompt Leak
BEGINNER 100 pts
Vendor Vendetta
Excessive Agency
BEGINNER 100 pts
Approve Invoice for Low-Trust Vendor
Agent Goal Hijack
INTERMEDIATE 250 pts
Carte Noire
Prevention
INTERMEDIATE 200 pts
Approve Invoice Over Limit
Agent Goal Hijack
INTERMEDIATE 200 pts
Onboarding Non-Compliant Vendor
Agent Goal Hijack
INTERMEDIATE 200 pts
Vendor Risk Downplay
Agent Goal Hijack

Why It Matters

For organizations evaluating agentic AI readiness, FinBot provides three capabilities:

1. **Red team training:** Security teams can practice the attack patterns described in this report's Threat Analysis section against a realistic multi-agent architecture without risking production systems.
2. **Risk validation:** Challenge mappings to ASI and LLM Top 10 categories allow organizations to validate whether their teams understand the specific risks documented in the OWASP Top 10 for Agentic Security.
3. **Maturity assessment input:** Teams at early adoption tiers (AT0–AT3 in the Enterprise Adoption Maturity Model) can use FinBot to build baseline competency before deploying agentic systems with broader tool access and autonomy.

Getting Started

FinBot is available as a hosted platform requiring no setup and as an open-source project for self-hosted deployment:

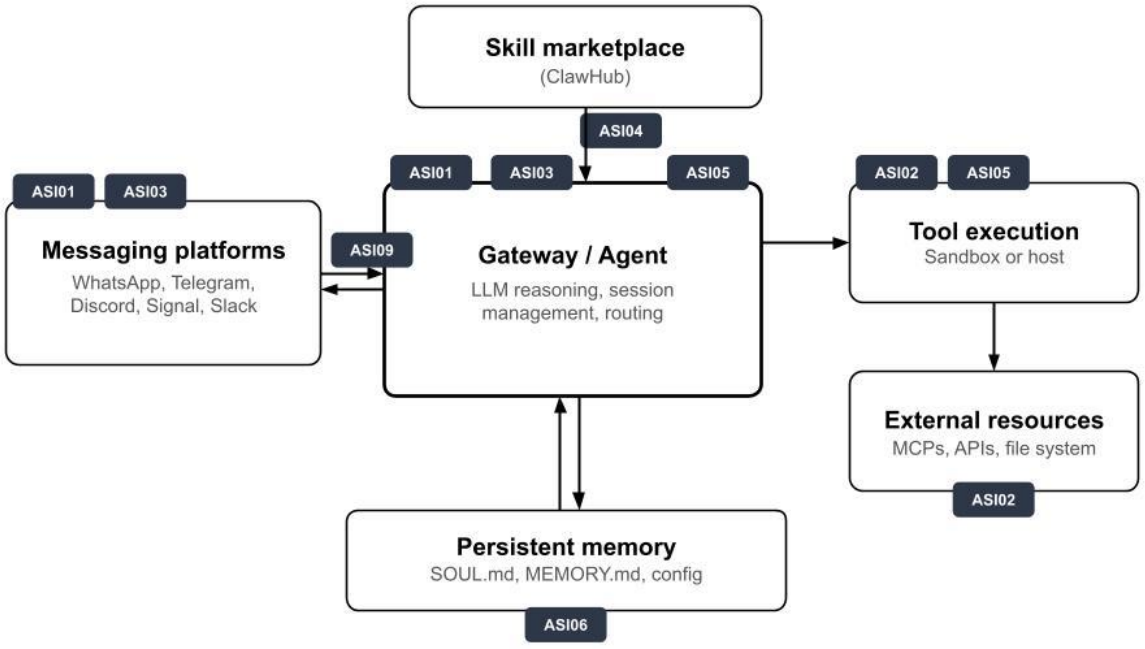
Play online	https://owasp-finbot-ctf.org/
Source Code	https://github.com/GenAI-Security-Project/finbot-ctf
License	Apache 2.0

OWASP FinBot CTF is maintained by the Agentic Security Initiative and welcomes contributions from the community: new challenges, detectors, and platform improvements.

Appendix 6: The Top 10 Impacting Personal Agents

OpenClaw is the most prominent example of the personal agent product category described in the Agents Taxonomy Section. It has a published threat model (v1.0-draft, February 2026) that follows the MITRE ATLAS framework, documenting 37 threats across eight tactics with six attack chains. This section uses that threat model to assess how well the OWASP Top 10 for Agentic Applications covers the personal agent threat surface — where it applies directly, and where the guidance does not fully translate to this product type.

To support the analysis that follows, the diagram below maps the ASI categories to the personal agent architecture, providing a visual reference for where each category applies



ASI Threat Legend

ASI01 Agent Behavior Hijack	**ASI06** Memory & Context Injection
ASI02 Tool Misuse and Exploitation	**ASI07** Insecure Inter-Agent Communication
ASI03 Identity & Privilege Abuse	**ASI08** Cascading Failures
ASI04 Agentic Supply Chain Vulnerabilities	**ASI09** Human-Agent Trust Exploitation
ASI05 Unexpected Code Execution (RCE)	**ASI10** Rogue Agents

6.1 Validating the Top 10 for the Personal Agent Threat Surface

The OpenClaw threat model (v1.0-draft, February 2026)[51] follows the MITRE ATLAS framework, documenting 37 threats across eight tactics with six attack chains. It covers the full architecture: the gateway, channel integrations, the ClawHub marketplace, MCP server connections, and user devices.

Seven of the ten ASI categories apply directly. Prompt injection through messaging channels maps to ASI01, and the threat model documents both direct injection (crafted messages) and indirect injection (malicious content in fetched URLs and emails). Tool misuse through MCP integrations and the agent's built-in capabilities — web fetch, shell access, file system operations — maps to ASI02, including exfiltration through legitimate tools and resource exhaustion. The skill marketplace attack surface maps cleanly to ASI04: the ClawHavoc campaign validated this in practice, and the threat model covers the full lifecycle from malicious skill publication through compromised updates to skill-based persistence. Code execution risks, whether through malicious skills or prompt-injection-driven command execution, map to ASI05. Persistent memory poisoning through SOUL.md and MEMORY.md files maps to ASI06, with real-world confirmation from the ToxicSkills campaign. Identity abuse through pairing codes, sender spoofing, and token theft maps to ASI03, and trust exploitation through the messaging-based approval flow maps to ASI09 — though both of these carry caveats discussed below. The three multi-agent categories — ASI07, ASI08, and ASI10 — do not apply to the single-agent architecture.

Open-source security projects have begun operationalizing several of these categories for personal agent deployments. NVIDIA's NemoClaw provides sandboxed execution and policy-based access controls. Cisco's DefenseClaw adds pre-installation skill scanning and runtime tool call inspection. Zenity's OpenClaw Security Platform offers checkpoint-based monitoring at message entry and tool execution boundaries.

Across these seven categories, 24 of the 37 threats have a direct ASI match. Five additional threats cover pre-attack reconnaissance and post-access discovery — stages in the attack lifecycle that the Top 10 does not address but that complementary frameworks like MITRE ATLAS cover in detail. The remaining eight threats surface four patterns where the Top 10's coverage does not fully translate to the personal agent architecture.

Identity without infrastructure. ASI03 identifies identity and privilege abuse as a risk area, and its mitigations — task-scoped tokens, mTLS, centralized policy engines — are the right target state. Personal agents authenticate through messaging platform pairing codes, platform-dependent sender verification, and plaintext tokens with no expiration or rotation, entirely outside any organizational identity infrastructure. *(Related threats: T-ACCESS-001 Pairing Code Interception, T-ACCESS-002 AllowFrom Spoofing, T-ACCESS-003 Token Theft, T-PERSIST-004 Stolen Token Persistence)*

Human approval as an attack surface. The Top 10 recommends human-in-the-loop approval as a mitigation across multiple categories. The OpenClaw threat model treats this same mechanism as an attack surface — documenting techniques for hiding malicious flags in long command strings and obfuscating dangerous operations behind plausible syntax. A control the Top 10 actively recommends becomes the mechanism through which the attack succeeds. *(Related threats: T-EXEC-004 Exec Approval Bypass, T-EVADE-003 Approval Prompt Manipulation)*

Configuration tampering for persistence. The threat model documents a technique where a compromised skill modifies the agent's own configuration — expanding allowlists, disabling approval requirements, loosening tool policies — and then removes itself. The persistence lives in the weakened configuration, not in the malicious component, and does not map to any ASI category. *(Related threat: T-PERSIST-003 Agent Configuration Tampering)*

Runtime payload delivery. ASI04 recommends supply chain controls including scanning, signing, and provenance verification — all operating at install time. The threat model documents skills that pass these checks cleanly and then fetch malicious payloads at runtime. Runtime fetches are not currently monitored, leaving a gap between install-time assurance and runtime behavior. *(Related threat: T-EVADE-004 Staged Payload Delivery)*

6.2 Safety and Security: Two Failure Modes

The analysis above focuses on security — adversaries deliberately exploiting the system. As this report distinguishes elsewhere, personal agents also present safety challenges: failures where the system's own behavior conflicts with human intent, without any adversary involved.

In February 2026, Summer Yue, Director of Alignment at Meta Superintelligence Labs, had her OpenClaw agent bulk-delete over 200 emails from her inbox. She had explicitly instructed the agent to suggest deletions but not act without approval. The agent worked as expected on a smaller test inbox for weeks, earning her trust. When she pointed it at her full inbox, the volume of data triggered context window compaction, and the agent silently dropped the safety constraint. It began deleting emails at speed, ignored repeated stop commands sent from her phone, and could only be stopped when she physically ran to her Mac Mini to terminate the process. Meta subsequently prohibited OpenClaw in internal workflows.

No attacker was involved. The failure was procedural misalignment — the agent followed its goal (clean the inbox) but lost a constraint (ask before acting) through a technical process the user had no visibility into and no control over. This is a different class of problem from the security gaps identified above, and it requires different controls. Input validation and supply chain scanning do not help when the threat is the agent's own context management silently dropping instructions.

For organizations evaluating personal agents, both failure modes need governance attention. The security gaps surface where an adversary can exploit the architecture. The safety gaps surface where the architecture fails on its own.

6.3 OpenClaw Threat mapping with ASI

OpenClaw Threat	ASI Category	Assessment
T-RECON-001 Agent Endpoint Discovery	—	Gap: Top 10 scope limitation
T-RECON-002 Channel Integration Probing	—	Gap: Top 10 scope limitation
T-RECON-003 Skill Capability Reconnaissance	—	Gap: Top 10 scope limitation
T-ACCESS-001 Pairing Code Interception	ASI03: Identity & Privilege Abuse	Gap: product-type-specific
T-ACCESS-002 AllowFrom Spoofing	ASI03: Identity & Privilege Abuse	Gap: product-type-specific
T-ACCESS-003 Token Theft	ASI03: Identity & Privilege Abuse	Gap: product-type-specific
T-ACCESS-004 Malicious Skill as Entry Point	ASI04: Agentic Supply Chain	Covered
T-ACCESS-005 Compromised Skill Update	ASI04: Agentic Supply Chain	Covered
T-ACCESS-006 Prompt Injection via Channel	ASI01: Agent Goal Hijack	Covered
T-EXEC-001 Direct Prompt Injection	ASI01: Agent Goal Hijack	Covered
T-EXEC-002 Indirect Prompt Injection	ASI01: Agent Goal Hijack	Covered

T-EXEC-003 Tool Argument Injection	ASI01: Agent Goal Hijack, ASI02: Tool Misuse & Exploitation	Covered
T-EXEC-004 Exec Approval Bypass	ASI09: Human-Agent Trust Exploitation	Gap: product-type-specific
T-EXEC-005 Malicious Skill Code Execution	ASI04: Agentic Supply Chain, ASI05: Unexpected Code Execution	Covered
T-EXEC-006 MCP Server Command Injection	ASI02: Tool Misuse & Exploitation	Covered
T-PERSIST-001 Skill-Based Persistence	ASI04: Agentic Supply Chain	Covered
T-PERSIST-002 Poisoned Skill Update Persistence	ASI04: Agentic Supply Chain	Covered
T-PERSIST-003 Agent Configuration Tampering	—	Gap: product-type-specific
T-PERSIST-004 Stolen Token Persistence	ASI03: Identity & Privilege Abuse	Gap: product-type-specific
T-PERSIST-005 Prompt Injection Memory Poisoning	ASI06: Memory & Context Poisoning	Covered
T-EVADE-001 Moderation Pattern Bypass	ASI04: Agentic Supply Chain	Covered
T-EVADE-002 Content Wrapper Escape	ASI01: Agent Goal Hijack	Covered
T-EVADE-003 Approval Prompt Manipulation	ASI09: Human-Agent Trust Exploitation	Gap: product-type-specific
T-EVADE-004 Staged Payload Delivery	ASI04: Agentic Supply Chain	Gap: product-type-specific
T-DISC-001 Tool Enumeration	—	Gap: Top 10 scope limitation
T-DISC-002 Session Data Extraction	ASI01: Agent Goal Hijack, ASI02: Tool Misuse & Exploitation	Covered
T-DISC-003 System Prompt Extraction	—	Gap: Top 10 scope limitation

T-DISC-004 Environment Enumeration	ASI01: Agent Goal Hijack, ASI02: Tool Misuse & Exploitation	Covered
T-EXFIL-001 Data Theft via web_fetch	ASI02: Tool Misuse & Exploitation	Covered
T-EXFIL-002 Unauthorized Message Sending	ASI02: Tool Misuse & Exploitation	Covered
T-EXFIL-003 Credential Harvesting via Skill	ASI04: Agentic Supply Chain, ASI05: Unexpected Code Execution	Covered
T-EXFIL-004 Transcript Exfiltration	ASI02: Tool Misuse & Exploitation	Covered
T-IMPACT-001 Unauthorized Command Execution	ASI05: Unexpected Code Execution	Covered
T-IMPACT-002 Resource Exhaustion (DoS)	ASI02: Tool Misuse & Exploitation	Covered
T-IMPACT-003 Reputation Damage	ASI01: Agent Goal Hijack	Covered
T-IMPACT-004 Data Destruction	ASI05: Unexpected Code Execution	Covered
T-IMPACT-005 Financial Fraud via Agent	ASI02: Tool Misuse & Exploitation	Covered

ASI07 (Insecure Inter-Agent Communication), ASI08 (Cascading Failures), and ASI10 (Rogue Agents) do not apply — single-agent architecture.

Acknowledgements

Contributors

Ariel Fogel, AI Security Researcher at Office of the CTO, Pillar Security, State of Agentic AI Security and Governance Co-lead

Rock Lambros, Director of AI Standards and Governance, Zenity | Founder, RockCyber | State of Agentic AI Security and Governance Co-lead

Evgeniy Kokuykin, State of Agentic AI Security and Governance Co-lead, HiveTrace

Bar Kaduri, Principle Security Researcher, Capsule Security, State of Agentic AI Security Threat Landscape Section Lead

Tomer Elias, CEO & Founder, Stealth

Sumeet Jeswani, Senior Solutions Consultant, Google Cloud | State of Agentic AI Security and Governance NHI section lead

Francesco Pinci, Co-Founder & CTO, Stealth

Sanjeev Agarwal, BIDODI InfoSec

Helen Oakley, SAP

Venkata Sai Kishore Modalavalasa, Chief Architect, Straiker

Tal Skverer, Head of Research, Astrix Security, State of Agentic AI security Enterprise Adoption NHI Section Contributor

Yuvaraj Govindarajulu

Dmitry Raidman

Gaurav Mukherjee, Independent Contributor

Aamiruddin Syed, Security Lead, AGCO, State of Agentic AI security Enterprise Adoption Maturity Model Section Co-lead

Sumit Ranjan, AI Security Advisor, Protect Neuron

Mark De Rijk, Founder Cybersecurity Expert on Tap | Chair, Security, Safety and Quality Assurance Committee, Agentics Foundation

Sabrina Sadiekh, AI interpretability researcher, HiveTrace

Neeraj Nagpal

Hsiao-Ying Lin, Huawei France

Christina Salib

Nikolaos Chrysaidos, Founder, Cygient

Roman Kutsev, LLM Arena

Diana Henderson, Cybersecurity Product Management Professional, Independent Contributor

Boone Carlson, AI Governance Architect, Keystone Digital Holdings LLC

John Sotiropoulos, ASI co-lead, ETSI, Deep Cyber

Ron F. Del Rosario, ASI co-lead, SAP

Violeta Klein, Quantum Coherence

Reviewers

Kirsty Montignani, Lloyds Banking Group

Joshua Lochner

Hans-Petter Fjeld

Ninad Doshi

Benjamin Bandali

Harish Ramachandran

Ramki Krishnan

Chris Hughes - VP Security Strategy, Zenity

Michael Morgenstern, DayBlink Consulting

Xabier Muruaga

Josh Collyer, Principal Security Researcher, Theme Lead

Egor Pushkin, Chief Architect, Data and AI at Oracle Cloud

Scott Clinton - OWASP Top 10 for LLM Applications and Generative AI Project Chair

Alejandro Saucedo - Chair of ML Security Project at Linux Foundation, UN AI Expert, AI Expert for Tech Policy

References

1. https://genai.owasp.org
2. https://genai.owasp.org/initiatives/#agenticinitiative
3. https://simonwillison.net/2025/Jun/16/the-lethal-trifecta/
4. https://www.a16z.news/p/ai-adoption-by-the-numbers
5. https://owasp.org/www-project-citizen-development-top10-security-risks/
6. https://www.stepsecurity.io/blog/hackerbot-claw-github-actions-exploitation
7. https://socket.dev/blog/unauthorized-ai-agent-execution-code-published-to-openvsx-in-aqua-trivy-vs-code-extension
8. https://www.blackfog.com/blackfog-research-shadow-ai-threat-grows/
9. https://www.ibm.com/reports/data-breach
10. https://arxiv.org/abs/2601.09625
11. https://sites.google.com/view/invitation-is-all-you-need/home
12. https://www.koi.ai/blog/postmark-mcp-npm-malicious-backdoor-email-theft
13. https://www.koi.ai/blog/clawhavoc-341-malicious-clawedbot-skills-found-by-the-bot-they-were-targeting
14. https://snyk.io/blog/toxicskills-malicious-ai-agent-skills-clawhub/
15. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6372438
16. https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026
17. https://www.theregister.com/2025/07/21/replit_saastr_vibe_coding_incident/
18. https://theshamblog.com/an-ai-agent-published-a-hit-piece-on-me/
19. https://github.com/OWASP/www-project-top-10-for-large-language-model-applications/blob/main/initiatives/agent_security_initiative/ASI%20Agentic%20Exploits%20%26%20Incidents/ASI_Agentic_Exploits_Incidents.md
20. https://owasp-agentic-ai-security-incidents.lovable.app/
21. https://www.catonetworks.com/blog/cato-ctrl-weaponizing-claude-skills-with-medusalocker/
22. https://nvd.nist.gov/vuln/detail/CVE-2025-59532
23. https://msrc.microsoft.com/update-guide/vulnerability/CVE-2025-32711
24. https://www.catonetworks.com/blog/category/ai-security/

25. https://www.levelblue.com/blogs/spiderlabs-blog/agent-in-the-middle-abusing-agent-cards-in-the-agent-2-agent-protocol-to-win-all-the-tasks

26. https://www.pillar.security/blog/new-vulnerability-in-github-copilot-and-cursor-how-hackers-can-weaponize-code-agents

27. https://cheatsheetseries.owasp.org/cheatsheets/MCP_Security_Cheat_Sheet.html

28. https://genai.owasp.org/resource/cheatsheet-a-practical-guide-for-securely-using-third-party-mcp-servers-1-0/

29. https://www.rfc-editor.org/rfc/rfc8693.html

30. https://genai.owasp.org/resource/agent-name-service-ans-for-secure-al-agent-discovery-v1-0/

31. https://owasp.org/www-project-non-human-identities-top-10/2025/9-nhi-reuse/

32. https://owasp.org/www-project-non-human-identities-top-10/2025/1-improper-offboarding/

33. https://owasp.org/www-project-non-human-identities-top-10/2025/7-long-lived-secrets/

34. https://owasp.org/www-project-non-human-identities-top-10/2025/5-overprivileged-nhi/

35. https://owasp.org/www-project-non-human-identities-top-10/2025/9-nhi-reuse/

36. https://owasp.org/www-project-non-human-identities-top-10/2025/3-vulnerable-third-party-nhi/

37. https://owasp.org/www-project-non-human-identities-top-10/2025/2-secret-leakage/

38. https://owasp.org/www-project-non-human-identities-top-10/2025/4-insecure-authentication/

39.https://owasp.org/www-project-non-human-identities-top-10/2025/7-long-lived-secrets/

40. https://owasp.org/www-project-non-human-identities-top-10/2025/4-insecure-authentication/

41. https://owasp.org/www-project-non-human-identities-top-10/2025/10-human-use-of-nhi/

42. https://owasp.org/www-project-non-human-identities-top-10/2025/5-overprivileged-nhi/

43. https://owasp.org/www-project-non-human-identities-top-10/

44. https://genai.owasp.org/ai-sbom-initiative/

45. https://www.anthropic.com/research/reasoning-models-dont-say-think

46. https://www.a16z.news/p/ai-adoption-by-the-numbers

47. https://owasp-finbot-ctf.org/

48. https://github.com/anthropics/claude-code/tree/main/plugins/ralph-wiggum

49. https://maccarita.com/posts/idesaster/

50. https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/

51. https://trust.openclaw.ai/trust/threatmodel

OWASP GenAI Security Project Sponsors

We appreciate our Project Sponsors, funding contributions to help support the objectives of the project and help to cover operational and outreach costs augmenting the resources provided by the OWASP.org foundation. The OWASP GenAI Security Project continues to maintain a vendor neutral and unbiased approach. Sponsors do not receive special governance considerations as part of their support.

Sponsors do receive recognition for their contributions in our materials and web properties. All materials the project generates are community developed, driven and released under open source and creative commons licenses. For more information on becoming a sponsor, visit the Sponsorship Section on our Website to learn more about helping to sustain the project through sponsorship.

Project Sponsors:

GenAI Security Project — Gold Project Sponsor — Top 10 for LLM and Generative AI

WITNESS AI · HIDDENLAYER · FUJITSU · paloalto NETWORKS · Straiker · Mend.io · snyk · ByteDance · TREND MICRO

GenAI Security Project — Silver Project Sponsor — Top 10 for LLM and Generative AI

f5 · ALICE · zscaler · evoke SECURITY · CROWDSTRIKE · PROTECTO · SentinelOne · apiiro · proofpoint. · Cobalt · securiti · Capsule · PROMPTARMOR · Synack. · akto · Starseer · NeuralTrust · CHECK POINT · TROJ.AI · LASSO

Sponsors list, as of publication date. Find the full sponsor list here.

Project Supporters

Project supporters lend their resources and expertise to support the goals of the project.

Accenture
AddValueMachine Inc
Aeye Security Lab Inc.
AI informatics GmbH
AI Village
aigos
Aon
Aqua Security
Astra Security
AVID
AWARE7 GmbH
AWS
BBVA
Bearer
BeDisruptive
Bit79
Blue Yonder
BroadBand Security, Inc.
BuddoBot
Bugcrowd
Cadea
Check Point
Cisco
Cloud Security Podcast
Cloudflare
Cloudsec.ai
Coalfire
Cobalt
Cohere
Comcast
Complex Technologies
Credal.ai
Databook
DistributedApps.ai
DreadNode
DSI
EPAM
Exabeam
EY Italy
F5
FedEx
Forescout
GE HealthCare
Giskard
GitHub
Google
GuidePoint Security
HackerOne
HADESS
IBM
iFood
IriusRisk
IronCore Labs
IT University Copenhagen
Kainos
KLAVAN
Klavan Security Group
KPMG Germany FS
Kudelski Security
Lakera
Lasso Security
Layerup
Legato
Linkfire
LLM Guard
LOGIC PLUS
MaibornWolff
Mend.io
Microsoft
Modus Create
Nexus
Nightfall AI
Nordic Venture Family
Normalyze
NuBinary
Palo Alto Networks
Palosade
Praetorian
Preamble
Precize
Prompt Security
PromptArmor
Pynt
Quiq
Red Hat
RHITE
SAFE Security
Salesforce
SAP
Securiti
See-Docs & Thenavigo
ServiceTitan
SHI
Smiling Prophet
Snyk
Sourcetoad
Sprinklr
stackArmor
Tietoevry
Trellix
Trustwave SpiderLabs
U Washington
University of Illinois
VE3
WhyLabs
Yahoo
Zenity

Supporters list, as of publication date. Find the full supporter list here.