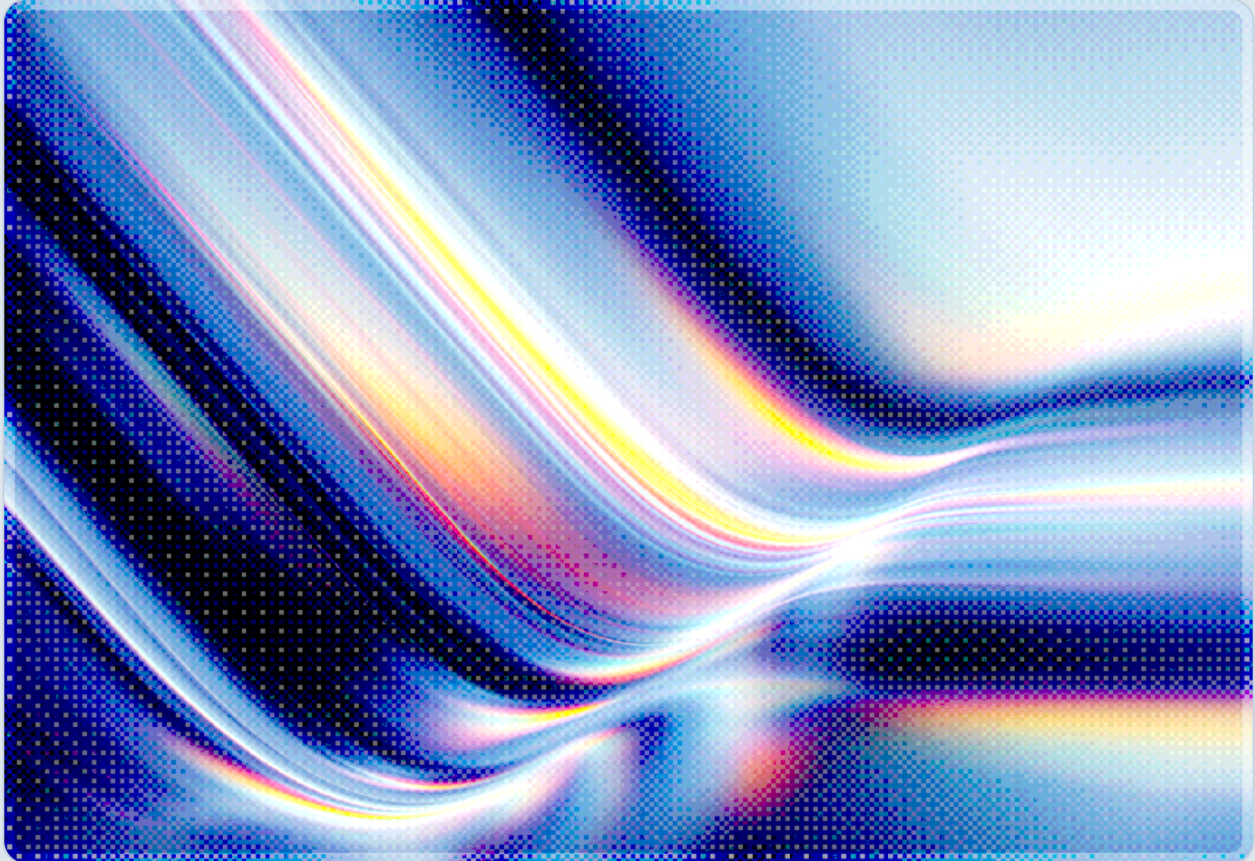


One Foundation, Ten Projects: Securing the AAIF Agent Stack

Security Governance Implications of the Agentic AI Foundation's
Consolidated Standards and Infrastructure

2026-10-29

 AI-assisted Rapid Research



CSAI

cloud
security
alliance®

© 2026 Cloud Security Alliance. Some rights reserved.

You may download, store, display, view, print, redistribute, and link to this document in its original, unmodified form, provided that attribution to the Cloud Security Alliance is maintained and all trademark and copyright notices remain intact.

This document may not be modified or altered. You may quote portions of the document as permitted by the Fair Use provisions of the United States Copyright Act, provided that attribution is given to the Cloud Security Alliance.

This document may be shared on professional and social media platforms in its original form with attribution.

This document was generated with AI assistance and has not undergone official CSA review and approval processes.

Key Takeaways

The Agentic AI Foundation (AAIF), a Linux Foundation body formed in December 2025, now hosts ten projects that together span most layers of a modern agent deployment: the protocols agents use to reach tools and each other, the file formats that carry their instructions and skills, the gateways that route their traffic, and the runtimes and sandboxes that execute their actions [1][2]. This consolidation simplifies interoperability, but it also concentrates a large share of the agent ecosystem's trust assumptions inside one governance structure. Security teams should treat the foundation's project portfolio as a dependency map, not as a certification of security maturity. Neutral governance is not a security baseline: AAIF's public project listing carries no maturity or security-readiness labels, and its hosted projects retain their own maintainers and release processes [5][7], so security posture varies by project.

Security maturity is also uneven across the specifications themselves. MCP's July 2026 revision tightened OAuth handling [10], and project leads reported at AGNTCon + MCPCon that roughly a third of its conformance scenarios now test security, a figure not yet backed by published data [3]. A2A v1.0 added signed agent cards [4], whereas the Agent Skills format defines no signing, integrity, or provenance fields [11]. This gap matters most for instruction-bearing artifacts. AGENTS.md files and Agent Skills package natural-language instructions and executable scripts that agents consume with significant privilege, neither format defines signing or provenance mechanisms [11], and public skill marketplaces have already shown measurable rates of critical issues [12].

Two further risks arise from how the projects fit together as of late October 2026. Gateways such as agentgateway and Agent Router route MCP, A2A, and model traffic together [6][7]; in CSA's assessment, a weakness in one specification or one gateway could therefore propagate across the stack. Protocol changes can also erode security tooling without warning. MCP's move to stateless operation removed protocol-level sessions [10], and monitoring or policy logic keyed on the session identifier may lose visibility after migration without raising errors.

Background

The Linux Foundation announced AAIF on December 9, 2025, anchored by three contributed projects: Anthropic's Model Context Protocol (MCP), Block's goose agent framework, and OpenAI's AGENTS.md convention for guiding coding agents [1]. The founding Platinum members were Amazon Web Services, Anthropic, Block, Bloomberg, Cloudflare, Google, Microsoft, and OpenAI [1]. At launch, the Linux Foundation reported more than 10,000 published MCP servers and adoption of AGENTS.md by more than 60,000 open source projects and agent frameworks, figures that indicate how widely these components were already embedded before they changed hands [1].

The foundation's governance took shape over the following months. By February 2026, AAIF reported 146 members, a Governing Board, and a Technical Committee composed of one representative from each Platinum member, which defines requirements for new projects and reviews proposals [9]. It also established seven working groups, among them Security and Privacy, Identity and Trust, and Governance, Risk and Regulatory Alignment [9]. The Technical Committee is currently chaired by Anthropic's David Soria Parra, MCP's co-creator and lead maintainer [8].

The portfolio then expanded steadily. agentgateway, originally donated to the Linux Foundation in 2025, was accepted as an AAIF project in 2026 [6]. Google's Agent2Agent (A2A) protocol, which had already moved to the Linux Foundation in 2025 and shipped its v1.0 specification in March 2026, joined AAIF as a hosted project on August 17, 2026 [4]. The Envoy AI Gateway left its position as a Cloud Native Computing Foundation (CNCF) Envoy sub-project and joined AAIF as Agent Router on September 10, 2026, with its code, maintainers, and Apache 2.0 license unchanged [5]. Finally, on October 22, 2026, at AGNTCon + MCPCon North America in San Jose, AAIF announced four further projects, namely Anthropic's Agent Skills, OpenAI's Open Responses, Hugging Face's OpenEnv, and Alibaba's OpenSandbox, bringing the total to ten [2]. Executive director Mazin Gilbert also announced three new governance committees at the event, although public coverage has not yet named them or described their remits [3].

The table below organizes the ten projects by the layer of the agent stack they occupy and summarizes the primary security surface each introduces.

Stack layer	Project (origin)	Joined AAIF	Primary security surface
Agent-to-tool protocol	Model Context Protocol (Anthropic)	Dec 2025	Tool authorization, OAuth flows, tool description integrity

Stack layer	Project (origin)	Joined AAIF	Primary security surface
Agent-to-agent protocol	Agent2Agent (Google)	Aug 2026	Agent identity, agent card authenticity, task delegation
Model interface	Open Responses (OpenAI)	Oct 2026	Cross-provider schema consistency, tool-call streaming
Instruction formats	AGENTS.md (OpenAI)	Dec 2025	Repository-borne instructions consumed by coding agents
Capability packaging	Agent Skills (Anthropic)	Oct 2026	Bundled instructions and executable scripts; marketplace supply chain
Traffic control plane	agentgateway (Solo.io)	2026*	AuthN/AuthZ, policy enforcement for MCP, A2A, and LLM traffic
Traffic control plane	Agent Router (Envoy AI Gateway, formerly a CNCF Envoy sub-project)	Sep 2026	Model and MCP access control, credential handling for backends
Agent runtime	goose (Block)	Dec 2025	Local code execution and file system access
Execution isolation	OpenSandbox (Alibaba)	Oct 2026	Sandbox lifecycle and escape resistance
Training and evaluation	OpenEnv (Hugging Face)	Oct 2026	Containerized environments exposing MCP and agent interfaces

Sources: AAIF and project announcements [1][2][4][5][6][7]. *The agentgateway documentation gives only the year of acceptance.

Security Analysis

Neutral Governance Is Not the Same as Security Assurance

The value of a neutral foundation is real. It reduces the risk that a single vendor changes a protocol unilaterally, it creates a public venue for specification changes, and it gives enterprises a stable counterparty for long-lived standards. Those benefits, however, concern *who decides*, not *how well* the resulting software resists attack. AAIF's public project page lists all ten projects without maturity, incubation, or security-readiness labels [7]. That differs from foundations such as the CNCF, whose graduation process requires a third-party security review and an OpenSSF Best Practices passing badge [18]. The foundation's February 2026 update described the Technical Committee as defining requirements for new projects, but it did not describe formal lifecycle stages, and project-specific governance models were still being written at that time [9]. CSA could not locate a published, foundation-wide security baseline (for example, a mandatory vulnerability disclosure policy, threat model, or third-party audit requirement) that applies uniformly to all hosted projects as of this writing.

The practical consequence is that an organization adopting "the AAIF stack" is in fact adopting ten artifacts with distinct maintainers, release cadences, and security histories. Agent Router's own announcement emphasizes that its maintainers and release process did not change on joining AAIF [5], which is a reasonable choice for continuity but confirms that foundation membership does not by itself alter a project's engineering practices. Security teams should therefore assess each component on its own record rather than inferring assurance from the foundation's brand.

Uneven Security Maturity Across the Specifications

The specifications themselves show markedly different levels of security engineering. MCP has arguably received the most public security scrutiny of the AAIF specifications, reflecting both its broad adoption and a series of early, severe implementation flaws. CVE-2025-6514 in the widely used mcp-remote client, for instance, allowed a malicious MCP server to achieve operating system command execution on the client by supplying a crafted OAuth `authorization_endpoint` value [13]. The MCP specification has since hardened its authorization model. The 2026-07-28 revision requires clients to validate the `iss` parameter, when the authorization server supplies one, against the recorded issuer before redeeming an authorization code (servers SHOULD include it, per RFC 9207). It also binds persisted client credentials to the authorization server that issued them and deprecates OAuth Dynamic Client Registration in favor of Client ID Metadata Documents [10]. Speakers at AGNTCon + MCPCon reported that roughly one third of MCP's 116 conformance scenarios now test security behavior, that enterprise-

managed authorization is marked stable on the authorization roadmap, and that proof-of-possession tokens and workload identity federation remain in draft [3]. InfoQ notes that several such figures were stage claims rather than published data, so they should be read as directional [3].

A2A has taken a different but comparable step: its v1.0 specification added signed agent cards to support cryptographic verification of an agent's identity and advertised capabilities [4]. By contrast, the Agent Skills specification defines a `SKILL.md` file with required `name` and `description` fields, optional `license`, `compatibility`, and `metadata` fields, and an experimental `allowed-tools` field, alongside an optional `scripts/` directory of executable code [11]. The specification contains no fields or guidance for signing, content hashing, publisher identity, or provenance, and its validation tooling checks frontmatter and naming conventions rather than behavior [11]. AGENTS.md is, by design, an even lighter convention. Neither format is a vulnerability in itself; the risk arises because both carry instructions an agent will act upon without any built-in means of verifying who authored them. Both now sit under the same governance umbrella as protocols whose recent revisions have added authentication and identity-verification features [4][10].

Instruction-Bearing Artifacts as a Supply Chain

The gap matters because skills and agent instruction files function as a software supply chain in which the "code" is partly natural language. Snyk's ToxicSkills study, published in February 2026, scanned 3,984 skills drawn from public skill marketplaces, principally ClawHub, and found that 534 of them (13.4%) carried at least one critical-severity issue, with 1,467 (36.8%) exhibiting some security issue and 76 malicious payloads confirmed through human review [12]. CSA's June 2026 research note on poisoned skills examined the same class of marketplace supply chain attack and concluded that skills warrant the vetting applied to software packages, covering their natural-language metadata as well as their code [17]. The MCP server ecosystem shows an analogous gap. OX Security's study of MCP registries, analyzed in a CSA research note in October 2026, found no consistent authorization model, operator identity, or lifecycle accountability across 15,465 listed MCP servers, and separate independent research summarized in the same note identified unverified lookalikes of branded servers [14].

Bringing Agent Skills into AAIF may improve this picture over time by giving the format a neutral venue in which integrity and provenance features could be standardized. That outcome is not yet assured. Until the specification addresses provenance, the portability that makes skills attractive, namely one format that works across agent products, also makes a malicious skill portable across those same products. Organizations should anticipate that attackers will target formats with broad cross-vendor reach, and foundation stewardship, if it accelerates cross-vendor adoption, would extend that reach further.

Composition Risk in the Control Plane

Two of the ten projects are gateways, and they occupy a position of unusual leverage. `agentgateway` describes itself as an agent-aware proxy that routes LLM traffic through an OpenAI-compatible API, supports MCP tool servers and A2A agent communication, and enforces JWT, API key, and MCP-spec authentication alongside role-based authorization written in the Common Expression Language (CEL) [6]. `Agent Router`, built on Envoy, connects applications to models and MCP tools with access and routing controls and manages backend credentials through dedicated security-policy resources [5][7]. Placing a gateway in front of agent traffic is a common recommendation, including in CSA's *Agentic MCP Security Best Practices Guide*. However, a gateway that terminates MCP, A2A, and model traffic also becomes a single point at which a parsing flaw, policy bypass, or credential leak can affect every protocol at once.

The same composition pattern appears elsewhere in the portfolio. `OpenEnv` environments expose an agent interface and MCP alongside the `Gym` training protocol [2], and `OpenSandbox` provides lifecycle and execution APIs that teams can extend with custom runtimes [2]. These overlaps are by design, since the projects are meant to interoperate, but they mean that a weakness in a shared component can affect every project that depends on it. A vulnerability class in MCP's schema handling, for example, could surface in a gateway, an evaluation environment, and a runtime simultaneously. CSA's view is that this interdependence argues for coordinated, cross-project vulnerability handling at the foundation level, which does not yet appear to be publicly documented.

Specification Churn Can Break Security Controls

The MCP 2026-07-28 revision also illustrates a quieter risk. It removed protocol-level sessions and the `Mcp-Session-Id` header, removed the initialization handshake in favor of per-request version and capability metadata, and replaced several server-initiated request types with a multi-round-trip pattern [10]. Some platforms reportedly already run about 70% of their MCP traffic in stateless mode, according to a conference presentation [3]. Security tooling that correlated activity by session identifier, or that inspected the initialization handshake to make policy decisions, may lose visibility after migration without raising an error. The revision does add standard `Mcp-Method` and `Mcp-Name` headers on HTTP requests and documents `OpenTelemetry` trace propagation conventions, both of which give defenders new hooks for policy and correlation [10]. MCP also adopted a formal deprecation policy with a minimum twelve-month window [10], which provides a predictable migration horizon. CSA did not identify equivalent published deprecation policies for the other AAIF specifications; adopting one would give adopters similar predictability.

Concentration of Influence

Finally, the consolidation raises a governance question that is not strictly technical. The Technical Committee consists of representatives from the eight Platinum members [8][9], several of which are also the original authors of projects with large installed bases, such as MCP and AGENTS.md [1], and major commercial vendors of agent products. This arrangement is common in industry foundations and is not evidence of misconduct. It does suggest that the security priorities of enterprise adopters, regulated industries, and smaller implementers may be underweighted unless those groups participate actively in the open working groups, which AAIF states are open to all membership tiers [9]. Competing interfaces also persist outside the foundation. Open Responses has drawn support from several platforms, while Anthropic and Google continue to use their own Messages and Interactions APIs respectively [3], so organizations should not assume that AAIF membership implies convergence on a single interface.

Recommendations

Immediate Actions

Organizations should first inventory which of the ten AAIF projects are present in their environment, including indirect use through vendor products, and record the specification revision each component implements. Particular attention should go to AGENTS.md files and Agent Skills, which developers may add to repositories and agent configurations without any security review. Teams should treat both as executable content by requiring code review for changes, restricting installation to an internal allowlist of approved skills, and scanning skill bundles for embedded scripts and instruction injection before use. Any remaining deployments of mcp-remote should be confirmed to run version 0.1.16 or later [13].

Short-Term Mitigations

Over the next one to three months, organizations should route MCP and A2A traffic through a policy-enforcing gateway with explicit authentication and per-tool authorization, while treating the gateway itself as a high-value asset that requires hardening, patching, and monitoring of its own. MCP clients should be upgraded to implementations that enforce the 2026-07-28 authorization requirements, including issuer validation and issuer-bound credentials [10], and teams should confirm that the authorization servers in use return the `iss` parameter. Where A2A is used, organizations should prefer signed agent cards and, where interoperability permits, reject unsigned cards from external parties, tracking exceptions explicitly [4]. Security monitoring that depended on MCP session identifiers should

be re-engineered around the new request headers and OpenTelemetry trace context [10]. Agent code execution, including goose and skill scripts, should run inside an isolation boundary with least-privilege file system and network access.

Strategic Considerations

Over the longer term, enterprises and their industry groups should engage with AAIF's Security and Privacy and Identity and Trust working groups to advocate for a foundation-wide security baseline covering a published threat model, a coordinated vulnerability disclosure process, security conformance tests, and a deprecation policy for every hosted project. A specific near-term ask is a signing, integrity, and provenance extension for Agent Skills comparable to A2A's signed agent cards. Procurement and third-party risk processes should ask vendors which AAIF specifications and revisions their products implement, how they handle cross-project vulnerabilities, and whether they pass the relevant conformance suites. Organizations should also plan for continued specification churn and budget for periodic re-validation of controls after each major protocol revision.

CSA Resource Alignment

CSA's **Agentic MCP Security Best Practices Guide**, currently in draft on CSA's labs site, is the most directly applicable prior work. It defines six MCP threat categories, seven control domains, and a four-level maturity model, and it recommends routing tool calls through an MCP gateway, sandboxing tool execution, enforcing OAuth 2.1 with short-lived scoped tokens, and verifying tool description integrity through hashing and signing. The findings in this note extend that guidance in two ways. First, the gateway and integrity controls it prescribes for MCP apply equally to A2A traffic and Agent Skills, which now share a governance home and, in many deployments, a gateway. Second, the MCP 2026-07-28 changes mean that some session-oriented monitoring practices should be revisited.

CSA's research note **15,465 MCP Servers and No Consistent Governance Signal: Marketplace Supply Chain Risk** documents the registry-level governance gaps that make instruction-bearing artifacts a supply chain concern [14]. Its recommendation to treat each MCP server as a governed third-party dependency, managed through an allowlist, short-lived credentials, and an egress gateway, maps directly onto the treatment this note recommends for Agent Skills and AGENTS.md content. Similarly, **Poisoned Skills: AI Agent Marketplace Supply Chain Attacks** addresses skill marketplaces directly [17]. Its guidance to vet skills as rigorously as software packages, require explicit approval and re-review after any change, grant agents minimal access, and monitor or sandbox skill behavior supports the allowlisting, scanning, and isolation recommendations above.

For threat modeling the consolidated stack, **MAESTRO**, CSA's agentic AI threat modeling framework, provides a layered approach well suited to the cross-layer composition risks described above, since the AAIF projects map onto distinct MAESTRO layers from foundation models and agent frameworks through deployment infrastructure and the agent ecosystem [15]. Control selection and assurance should be anchored in the **AI Controls Matrix (AICM) v1.1**, whose supply chain, identity and access management, and threat and vulnerability management domains provide auditable control objectives for the inventory, allowlisting, gateway, and vulnerability handling recommendations in this note [16].

References

- [1] Linux Foundation. "[Linux Foundation Announces the Formation of the Agentic AI Foundation \(AAIF\), Anchored by New Project Contributions Including Model Context Protocol \(MCP\),_goose and AGENTS.m](#)." Linux Foundation, December 9, 2025.
- [2] Linux Foundation. "[Agentic AI Foundation Announces Four New Projects: Agent Skills, Open Responses, OpenEnv, and OpenSandbox](#)." Linux Foundation, October 22, 2026.
- [3] Hoblitzell, Andrew. "[AGNTCon + MCPCon 2026: Agentic AI Foundation Adds Four Projects as MCP Turns to Security](#)." InfoQ, October 26, 2026.
- [4] Agentic AI Foundation. "[A2A Joins AAIF](#)." AAIF Blog, August 17, 2026.
- [5] Agent Router Maintainers. "[Envoy AI Gateway Is Becoming Agent Router, an Agentic AI Foundation Project](#)." Agent Router Blog, September 9, 2026.
- [6] agentgateway Project. "[agentgateway Documentation](#)." agentgateway.dev, accessed October 29, 2026.
- [7] Agentic AI Foundation. "[Projects](#)." AAIF, accessed October 29, 2026.
- [8] Agentic AI Foundation. "[Technical Committee](#)." AAIF, accessed October 29, 2026.
- [9] Agentic AI Foundation. "[AAIF's First Quarter Success Story: New Members, Technical Wins, and Open Governance](#)." AAIF Blog, February 24, 2026.
- [10] Model Context Protocol. "[Key Changes: Specification Revision 2026-07-28](#)." modelcontextprotocol.io, July 28, 2026.
- [11] Agent Skills. "[Agent Skills Specification](#)." agentskills.io, accessed October 29, 2026.
- [12] Snyk. "[ToxicSkills: Malicious AI Agent Skills on ClawHub](#)." Snyk Blog, February 5, 2026.
- [13] JFrog Security Research. "[Critical RCE Vulnerability in mcp-remote: CVE-2025-6514 Threatens LLM Clients](#)." JFrog Blog, July 9, 2025.
- [14] Cloud Security Alliance. "[15,465 MCP Servers and No Consistent Governance Signal: Marketplace Supply Chain Risk](#)." CSA AI Safety Initiative, October 7, 2026.

- [15] Cloud Security Alliance. "[Agentic AI Threat Modeling Framework: MAESTRO](#)." CSA Blog, February 6, 2025.
- [16] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, accessed October 29, 2026.
- [17] Cloud Security Alliance. "[Poisoned Skills: AI Agent Marketplace Supply Chain Attacks](#)." CSA AI Safety Initiative, June 24, 2026.
- [18] Cloud Native Computing Foundation. "[Project Graduation Application Template](#)." CNCF Technical Oversight Committee, GitHub, accessed October 29, 2026.