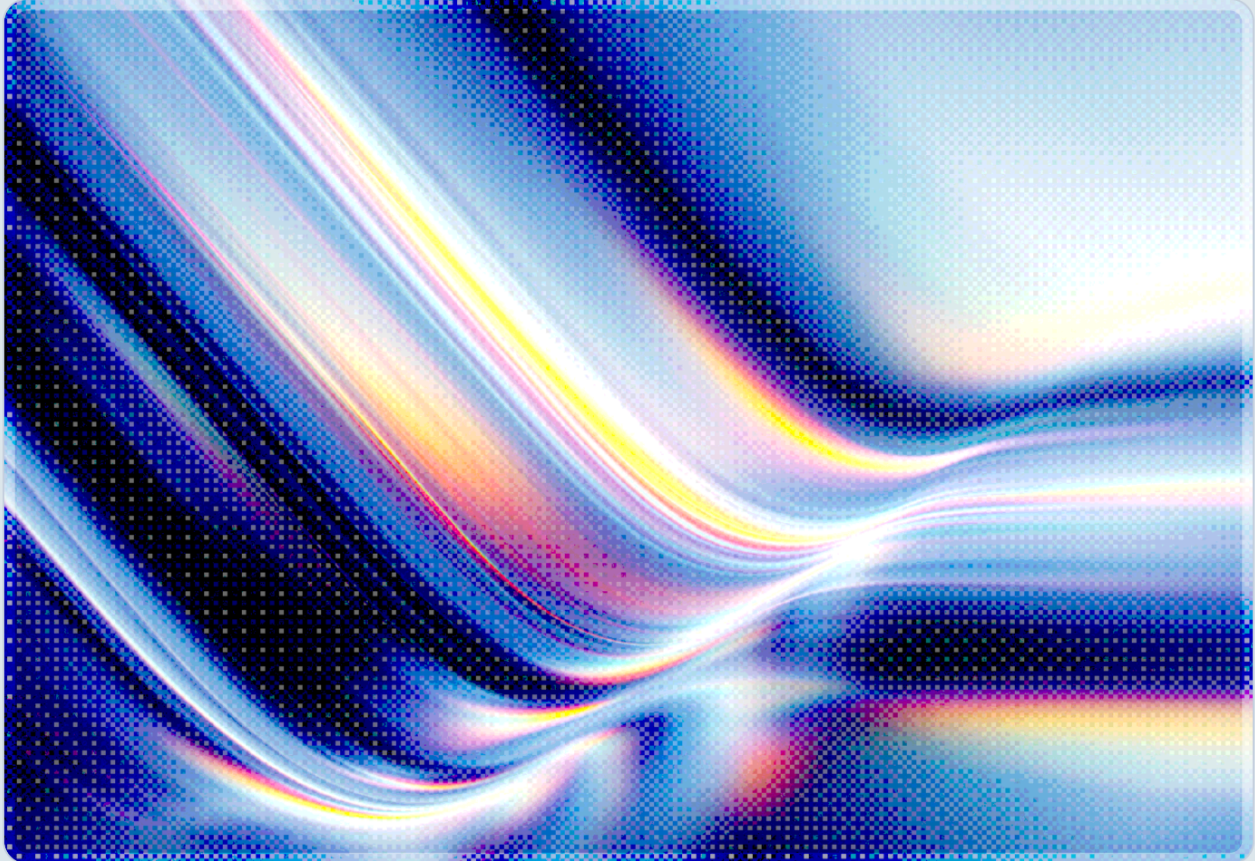


Who Is Defining Trustworthy Agentic AI?

A Global Landscape of AI and Agentic AI Standards, Frameworks, Protocols, and Assurance (as of October 27, 2026)

2026-10-27

 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.

Table of Contents

Executive Summary 4

1. Introduction and Background 7

2. Methods and Definitions 8

 2.1 Artifact definitions

 2.2 Organization types

 2.3 Evidence rules and limitations

 2.4 Selection rule for profiles and the top ten

 2.5 Conflict-of-interest policy

3. Market Map: Who Is Doing the Work 10

 3.1 The shape of the field

 3.2 Safe AI Society and other comparators

4. Artifact Map: Many Layers, Not One Winner 13

 4.1 The trust and standards stack

 4.2 Artifact feature matrix

5. Detailed Profiles 17

 5.1 ISO/IEC JTC 1 (SC 42, SC 27, SC 40)

 5.2 European Union: AI Act, AI Office and CEN-CENELEC JTC 21

 5.3 NIST and the Center for AI Standards and Innovation (CAISI)

 5.4 ETSI Technical Committee Securing AI (TC SAI)

 5.5 ITU-T (Study Groups 17, 20 and 21)

 5.6 IETF (OAuth and WIMSE working groups)

 5.7 OpenID Foundation, FIDO Alliance and W3C (identity cluster)

 5.8 Linux Foundation Agentic AI Foundation (MCP, A2A, AGENTS.md)

 5.9 OWASP GenAI Security Project

 5.10 Coalition for Secure AI (CoSAI), an OASIS Open Project

 5.11 MITRE: ATLAS, AI incident sharing, and the CWE/CVE ecosystem

 5.12 Cloud Security Alliance and CSAI Foundation (self-assessed)

 5.13 AIUC-1 (Artificial Intelligence Underwriting Company)

 5.14 Financial-sector baselines: U.S. Treasury and CRI, MAS, HKMA, FINOS

 5.15 Singapore: IMDA, AI Verify Foundation and the Cyber Security Agency

 5.16 United Kingdom: DSIT, NCSC and the AI Security Institute

 5.17 China: TC260, CAC and CAICT

 5.18 Japan, Korea, India and Australia

 5.19 IEEE Standards Association

 5.20 MLCommons and independent evaluators (METR, Apollo Research, FAR.AI)

 5.21 Frontier Model Forum and frontier developer frameworks

 5.22 Vendor and vendor-originated frameworks

 5.23 Software supply chain: OpenSSF, SLSA, Akrites, CIS and SAFECode

Editorial Status: Publication Hold 25

References 26

Executive Summary

Enterprises adopting AI agents in 2026 face a crowded and fast-moving field of guidance. Formal standards bodies, open-source foundations, industry coalitions, governments, frontier developers, auditors and advocacy groups all publish material described as a "standard" or "framework," yet these artifacts do very different jobs. This paper surveys that field as of October 27, 2026. It separates *who* does the work from *what kind* of artifact they produce and *how much independent evidence* exists that the artifact is used, and it then asks where the Cloud Security Alliance (CSA) overlaps, complements or could be displaced by others. The independent landscape study comes first; the CSA strategy chapter (Section 11) is explicitly labeled as strategic interpretation and carries a conflict-of-interest notice, because CSA is both the author and one of the subjects.

The central finding supports, with qualifications, the thesis the commissioning brief asked us to test rather than assume: the sector is settling into **layers rather than a single winning framework**. Management-system standards (ISO/IEC 42001), operational control catalogs (CSA AICM, AIUC-1, the U.S. Treasury Financial Services AI Risk Management Framework), threat taxonomies (MITRE ATLAS, OWASP), runtime and interoperability protocols (MCP, A2A, OAuth and WIMSE drafts), evaluation benchmarks (MLCommons AILuminate, UK AISI Inspect) and regulatory obligations (the EU AI Act and its first harmonised standard, China's TC260 and CAC instruments) each solve a different problem. In the cases examined, mappings between artifacts were more common than outright displacement. For example, CSA's AI Controls Matrix (AICM) v1.1 maps to AIUC-1 [1], CSA's STAR Registry now recognizes AIUC-1 certification [5], and OWASP has published an AIUC-1 crosswalk for its agentic Top 10 [28]. Two of these three examples are CSA's own actions, however. Mapping is also often one-directional, since AIUC-1 does not publish an AICM crosswalk, and the proliferation of catalogs at the operational layer remains a burden for buyers. Competition is nevertheless real at the operational control and assurance layer, where several catalogs (AICM, AIUC-1, Vanta's Agentic Trust Controls, Pillar's SAIL, the FS AI RMF) address overlapping buyer needs.

Ten consequential standard-setting ecosystems (selection rule in Section 2.4). These are ranked by durable governance rights, regulatory referencing, implementer base and assurance channel, not by familiarity. They are ISO/IEC JTC 1 (42001 and its accreditation chain, plus the newly issued ISO/IEC 27090); the EU AI Act with CEN-CENELEC JTC 21; NIST and its Center for AI Standards and Innovation (CAISI); the Linux Foundation's Agentic AI Foundation (MCP, A2A, AGENTS.md); the IETF and OpenID Foundation identity stack; the OWASP GenAI Security Project; MITRE (ATLAS, CWE, and its role in the CVE ecosystem); CSA (AICM and STAR for AI; *self-assessed*); the AIUC-1 standard and certification; and China's TC260 and Cyberspace Administration (CAC) with CAICT's ITU-T contributions. ETSI TC SAI, CoSAI, Singapore's IMDA and CSA Singapore, and ITU-T form a close next tier.

What changed in 2025–26. The first agent-specific artifacts moved from blog posts to governed processes. ITU-T approved Recommendation F.748.93 on AI agent interoperability on August 29, 2026 [52]. ETSI adopted a work item for baseline agentic AI cybersecurity requirements (TS 104 328) on October 1, 2026 [44] **[VERIFICATION NEEDED: no public ETSI work-programme record located]**. The IETF WIMSE working group adopted an AI Identity Management System draft in September 2026 [56]. NIST launched an AI Agent Standards Initiative in February 2026 [76]. A2A reached v1.0 and joined MCP under the Agentic AI Foundation [69] [70]. The first EU AI Act harmonised standard, EN 18286 on quality management, was cited in the Official Journal in October 2026 [46] [47], while the Digital Omnibus, Regulation (EU) 2026/1744 (published in the Official Journal on July 24, 2026), pushed high-risk obligations to December 2027 and August 2028 [49]. Assurance also changed: AIUC-1 introduced quarterly-updated, agent-level certification with accredited auditors [28][29], CSA launched STAR for AI Level 2 and Valid-AI-ted for AI [4], and national accreditation bodies began accrediting ISO/IEC 42001 certification bodies under ISO/IEC 42006:2025 [38][121][122].

Five most material coverage gaps (hypotheses tested in Section 10). The landscape review points to five gaps that matter most for enterprises delegating authority to agents. They are framed here as working hypotheses, and Section 10 sets out the test applied to each. The table below states each hypothesis alongside the evidence observed at the cutoff.

Gap hypothesis	Evidence at the cutoff
No ratified standard governs agent delegation chains end to end.	The IETF, CoSAI ODIS, Agentic Trust Framework and ITU-T Focus Group work all remain drafts, work items or individual submissions.

Gap hypothesis	Evidence at the cutoff
No runtime assurance baseline yet combines a published specification, multi-vendor governance and independent conformance testing.	Candidates include the OWASP Agent Control Standard (v0.1.0), CSA's AARM (v1.0), ETSI TS 104 328 (work item) and China's mandatory agent standard (in drafting).
Most certificates attest an organization or management system at a point in time, not a specific, versioned agent.	AIUC-1's quarterly testing is the only recurring agent-level technical test found.
The threat → control → evidence → response pipeline lacks a common machine-readable record.	ATLAS, AIVSS, CWE and CVE pilots are emerging but not joined up.
Digital-worker accountability and the evaluation of cyber-capable defensive agents lack assessable controls.	Both are addressed by guidance and research, not by assessable controls with independent verification.

Taken together, these gaps sit in the middle of the agentic trust stack described in Section 4, where identity, runtime control and evidence meet. That is also where formal standardization has made the least progress.

Census (deduplicated, verified entries; full inventory in Appendix A). The 94 seed lines in the commissioning brief mixed legal organizations, programs, working groups and comparators. After deduplication and verification, the census distinguishes the following categories.

Count category	Number	Basis
Unique active publishers or hosts of AI/agentic standards, frameworks, specifications or official guidance	62	Appendix A, census code P
Assurance, evaluation, accreditation and credential operators that do not own the underlying standard	15	Appendix A, census code A
Watch-list or non-standard organizations (advocacy, convening, unverified or no AI-specific deliverable)	16	Appendix A, census code W
Distinct artifacts tracked (standards, frameworks, taxonomies, protocols, benchmarks, schemes, laws)	92	Appendix B
– of which published or in force	69	Appendix B, status published/in force/active
– of which drafts, work items or under study	23	Appendix B, status draft/work item
Live third-party certification or attestation schemes for AI organizations, systems or agents	6	Section 8, Table 8.1

These counts describe this paper's verified census at a single date. They are not a market-share measure and will change quickly.

Adoption and assurance story. Verified adoption evidence is uneven. Protocol adoption (MCP, A2A) and taxonomy adoption (OWASP, ATLAS) is broad but mostly visible through implementations, membership counts and downloads, not audits. Assurance adoption is narrower but more verifiable: accredited ISO/IEC 42001 certification bodies exist in the United States and United Kingdom [120][121], AIUC-1 publicly lists certified companies [28], and the STAR Registry publishes entries [3]. No verifiable public count of ISO/IEC 42001 certificates or STAR for AI entries was found, so this paper does not estimate one.

Implications for boards and CISOs. Boards should not expect a single certificate to answer the question "is this agent safe to delegate authority to?" A defensible posture in 2026 combines a management-system certification (ISO/IEC 42001) for governance, a mapped operational control catalog for implementation and audit (such as AICM, AIUC-1 or a sector profile; CSA's interest in AICM is disclosed in Section 2.5), recurring independent testing for specific agents, and protocol-level identity and delegation controls that are, today, still drafts. CISOs should ask vendors which *version* of which agent, under which *scope*, was tested *when*, and what change would invalidate the claim. Section 8 provides a checklist for that conversation.

1. Introduction and Background

On October 27, 2026, a CSA board member asked for a competitive analysis of "what groups are working on what in AI frameworks in different categories," citing Safe AI Society as an example of a recently noticed nonprofit. The question is harder than it appears. "Framework" has become an elastic term, and many organizations with the word "safe" or "trust" in their names publish no implementable requirements at all. Conversely, some of the most consequential agent-security work is published as open-source protocol specifications or Internet-Drafts that never call themselves standards. A useful answer therefore needs a vocabulary that separates kinds of work before it compares organizations.

The commissioning brief posed two linked questions. The first is a landscape question: which organizations develop, convene, sponsor, publish, test, adopt, govern or enforce AI and agentic AI standards and frameworks? The second is a positioning question: which of that work overlaps CSA's current role, which complements it, which could displace it, which can be mapped or integrated, and which offers durable leadership opportunities? This paper answers the first question in Sections 3 through 10 using public sources only. It answers the second in Section 11, which is labeled throughout as CSA's strategic interpretation.

The analytical lens is enterprise adoption of AI and AI agents rather than AI ethics in general. The paper therefore gives greatest depth to security and safety governance, auditable controls, agent autonomy and delegated authority, identity, interoperability, runtime control, evaluation, resilience, and AI-enabled cyber defense. Organizations whose work is primarily ethical principles or public education are included in the census and discussed where relevant, but they are not profiled at equal depth.

Finally, a note on terminology. On October 20, 2026, NIST renamed its SP 800-53 Control Overlays for Securing AI Systems project to "Control Overlays for Super Intelligence Systems (COSIS)," citing a September 29, 2026 executive order directing U.S. executive agencies to use "Super Intelligence" in place of "Artificial Intelligence" [79]. NIST notes that previously issued publications were not altered [79]. This paper uses "AI" throughout and refers to U.S. federal artifacts by their published titles.

2. Methods and Definitions

2.1 Artifact definitions

The paper classifies every artifact using five types drawn from the commissioning brief. A **formal standard** has been approved through an identifiable standards development process (ISO/IEC, CEN-CENELEC, ETSI, ITU-T, IEEE, or an IETF RFC) with a status and version. An IETF Internet-Draft, an ISO Approved Work Item, an ETSI work item or an ITU-T item "under study" is *not* a formal standard. A **framework or control catalog** provides actionable domains, requirements, control objectives, and implementation or maturity guidance. A **risk or threat taxonomy** enumerates threats, tactics, weaknesses or risks and is not synonymous with auditable controls. A **protocol or specification** defines technical rules for interoperability, identity, permission, delegation, inter-agent interaction or evidence transport. A **testing or evaluation benchmark** provides reproducible assays, test suites or scoring methods. Two further types were needed in practice: **assurance schemes** (certification, attestation or registry programs) and **binding or official regulatory instruments**.

2.2 Organization types

Each organization is classified by primary role, with secondary roles noted in Appendix A. The types are formal standards development organization (SDO); open foundation, consortium or nonprofit publisher; government or intergovernmental body; commercial vendor or frontier developer; assurance, accreditation or audit provider; evaluation or research organization; and advocacy or convening body. A government regulator that also publishes technical guidance is classified as government; an SDO that also runs a conformity program is classified as an SDO.

2.3 Evidence rules and limitations

The evidence cutoff is October 27, 2026. The census started from the 94-item seed inventory in the commissioning brief, which was validated, expanded and deduplicated. Programs, working groups and projects were folded into their host organization for the organization count and tracked separately as artifacts. Substantive claims are cited to a public source, with version and date where available. Primary sources were preferred; where only secondary sources (law firms, trade press) were available, the text says so or marks the claim **[VERIFICATION NEEDED]**. Several primary pages (iso.org, eur-lex.europa.eu, tc260.org.cn) could not be retrieved directly during research, and claims depending on them are marked with medium confidence.

The paper avoids invented quantities. It reports adoption figures only where a publisher states them, attributes them, and does not convert them into market share. It distinguishes published artifacts from drafts, open-source implementations and reference architectures. Geographic balance was sought deliberately. The census includes Chinese, Singaporean, Japanese, Korean, Indian, Australian, Hong Kong, ASEAN and African Union instruments as well as U.S., EU and UK work, although depth on non-English primary texts is lower and is flagged.

This is a public-source document. It contains no private correspondence, Chatham House disclosures, unannounced deal terms or private roadmap claims. A separate internal annex, which is not part of this paper, addresses non-public CSA considerations.

2.4 Selection rule for profiles and the top ten

Deep-dive profiles (Section 5) were selected using a rule rather than familiarity. An organization or initiative was profiled if it met at least two of four criteria. The first is **governance durability**: it holds formal standardization rights or stewards a protocol with multi-vendor governance. The second is **regulatory referencing**: a regulator or official body cites, harmonises or mandates it. The third is **verified implementation or assurance**: it has public implementations, accredited auditors or registry entries. The fourth is **direct agentic scope**: it publishes agent-specific requirements, protocols or tests. The "top ten" in the Executive Summary are the ecosystems that scored highest on the standards-power assessment in Section 9.4, cross-checked against the scorecard in Appendix C.

2.5 Conflict-of-interest policy

CSA is the author of this paper and a subject of it. To manage this, CSA artifacts were scored with the same evidence bar as peers. CSA's self-scores are marked "self-assessed" in every table. CSA-authored research notes that state they were AI-assisted and not formally reviewed (several CSAI Foundation and CSA Labs notes) are cited as background, not as official CSA standards. The strategy chapter (Section 11) includes counterarguments and identifies areas where CSA is *not* uniquely positioned.

3. Market Map: Who Is Doing the Work

3.1 The shape of the field

The seed inventory treated organizations, programs and working groups as peers. Once these are deduplicated, a clearer structure appears. A small number of *hosts* sit at the center: ISO/IEC, the Linux Foundation, OASIS, OWASP, the IETF, ETSI, ITU, IEEE and CSA. Each hosts many initiatives, and several initiatives that look independent are projects of the same host. MCP, A2A, AGENTS.md, goose, SPIFFE/SPIRE, OpenSSF, FINOS, Akrites and the TRACE agent-attestation project all sit under the Linux Foundation umbrella [68][69] [73][90]. CoSAI and its ODIS delegation work are projects of OASIS Open [24][27]. Around these hosts sit governments that reference, mandate or fund standards, and vendors that originate many of the specifications the hosts later steward.

The market can be read along two axes. The first is publisher type, which drives neutrality, governance rights and the likelihood of regulatory referencing. The second is lane: which problem the artifact addresses, from management systems (L1) through public-interest advocacy (L13). Table 3.1 summarizes the deduplicated organizations by publisher type and geography, and Table 3.2 shows lane coverage. Full detail is in Appendix A.

Table 3.1: Organizations by publisher type and principal geography (deduplicated; counts from Appendix A)

Publisher type	Examples (not exhaustive)	Principal geographies	Typical artifact
Formal SDO (international, regional, national)	ISO/IEC JTC 1, ETSI, CEN-CENELEC, ITU-T, IETF, IEEE SA, W3C, China TC260	Global, EU, China	Standards, technical specifications, recommendations, RFCs
Open foundation, consortium or nonprofit publisher	CSA/CSAI, OWASP, OASIS/CoSAI, Linux Foundation/AAIF, OpenID Foundation, FIDO Alliance, MITRE (as operator of public knowledge bases), MLCommons, SAFECODE, ISACA, Frontier Model Forum	Global, U.S.-weighted	Control catalogs, taxonomies, protocols, benchmarks, credentials
Government or intergovernmental	NIST/CAISI, EU Commission/AI Office, ENISA, UK DSIT and AISI, Singapore IMDA, CSA Singapore and MAS, HKMA, Germany BSI, France ANSSI, China CAC, Japan, Korea, India, Australia, OECD, UNESCO, Council of Europe, ASEAN	All regions	Law, official guidance, evaluation tools, profiles
Commercial vendor or frontier developer	AIUC (AIUC-1), Google, Microsoft, NVIDIA, AWS, Cisco, Anthropic, OpenAI, Meta, Vanta, Pillar Security	U.S.-weighted	Vendor frameworks, frontier safety policies, open-source runtimes, commercial standards
Assurance, accreditation or audit provider	ANAB, UKAS, Schellman, Coalfire, BSI Group, PwC and other audit firms, AICPA (criteria owner)	U.S., UK, global	Certificates, attestations, accreditation
Evaluation or research organization	METR, Apollo Research, FAR.AI, Center for AI Safety, Canada CAISI	U.S., Canada, UK	Evaluations, benchmarks, research methods

Publisher type	Examples (not exhaustive)	Principal geographies	Typical artifact
Advocacy, convening or watch-list	Safe AI Society, Partnership on AI, Future of Life Institute, WEF, AI Alliance, Concordia AI, UN dialogue mechanisms	Global	Principles, indices, convenings, reports

The geography column reveals a structural imbalance that matters for the strategy discussion. Implementable agent-security artifacts are concentrated in U.S.-headquartered foundations and vendors and in European SDOs. China operates a parallel, state-coordinated stack that is increasingly agent-specific and is exported partly through ITU-T. Singapore, Japan and Korea publish government frameworks and evaluation guidance that are often more agent-specific than their Western counterparts, but they rely on ISO/IEC and their own institutes rather than private consortia.

Table 3.2: Lane coverage by representative initiatives (L1-L13)

Lane	Question	Most active publishers (verified)	Maturity in 2026
L1 Governance and management systems	Is AI governed as a managed system?	ISO/IEC 42001/42005/42006, NIST AI RMF, CEN-CENELEC EN 18286, Singapore MGF	Mature; certifiable
L2 Control catalogs and frameworks	What controls apply and who owns them?	CSA AICM v1.1, AIUC-1, Treasury/CRI FS AI RMF, ETSI EN 304 223, BSI AIC4, FINOS AIGF, Vanta ATC, Pillar SAIL	Mature but crowded
L3 Risk and threat taxonomies, secure development	What can go wrong?	MITRE ATLAS, OWASP LLM and Agentic Top 10, NIST AI 100-2, CoSAI Risk Map, Microsoft failure-mode taxonomy, ISO/IEC 27090	Mature; agent coverage growing monthly
L4 Agent identity, delegated authority, privileges	Who authorized this agent, and can authority be revoked?	IETF OAuth and WIMSE, OpenID Foundation, CoSAI ODIS, FIDO Alliance, ITU-T Focus Group, ATF, NIST NCCoE	Mostly drafts
L5 Protocols, orchestration, interoperability	How do agents, tools and services interact?	AAIF (MCP, A2A, AGENTS.md), ITU-T F.748.93, IEEE P3394, W3C community groups, Eclipse LMOS	Fast-moving; de facto standards ahead of formal ones
L6 Runtime controls, isolation, kill switches	How are live actions constrained and halted?	OWASP ACS, CSA AARM, NVIDIA OpenShell, ETSI TS 104 328 (work item), China agent standard (drafting)	Early
L7 Evaluations, safety testing, red teaming	How is effectiveness measured?	UK AISI Inspect, MLCommons, METR, Apollo, Japan AISI, AI Verify, ISO/IEC 42119 series	Active research; few accredited schemes
L8 Audit, certification, registries	Who checks what, and can customers verify?	ISO/IEC 42001 certification (ANAB, UKAS), CSA STAR for AI, AIUC-1, BSI AIC4/C5, IEEE CertifAIEd	Emerging market
L9 Software and supply chain, AI-generated code	How is machine-written code and the model supply chain secured?	OpenSSF, SLISA, CoSAI WS1, CIS/SAFECode, Akrites	Moderate; AI-specific guidance thin

Lane	Question	Most active publishers (verified)	Maturity in 2026
L10 Frontier and catastrophic risk	How are high-consequence capabilities controlled?	Frontier Model Forum, Anthropic RSP, OpenAI Preparedness, Google DeepMind FSF, EU GPAI Code, CSA Catastrophic Risk Annex (work item)	Corporate policies; regulatory code in EU
L11 Sector, regional, regulatory requirements	How do obligations differ by jurisdiction and sector?	EU AI Act, China CAC/TC260, Korea AI Basic Act, MAS, HKMA, Treasury, FDA, ISO/PAS 8800	Diverging timelines
L12 AI-powered cyber defense, machine-speed operations	How are defensive agents assessed and governed?	NIST IR 8596 (draft), UK AISI Inspect Cyber, FMF "Agents for Cyber Defense," Akrites	Thin; mostly research
L13 Advocacy, coordination, incubators	Who shapes norms without implementable standards?	OECD, UNESCO, UN Global Dialogue, WEF, PAI, FLI, Safe AI Society	Active; non-normative

3.2 Safe AI Society and other comparators

The board question used Safe AI Society as an example, so it warrants a direct answer. Safe AI Society describes itself as a grassroots nonprofit that educates the public about AI progress and risk and seeks to influence experts and policymakers, principally through a newsletter [35]. Its site names no leadership and shows no published technical controls, standardization work, audit criteria or assurance program [35]. It is therefore classified as an advocacy organization on the watch list (lane L13), not as a technical competitor to CSA. It should not be confused with the Safe AI Forum, which runs the International Dialogues on AI Safety and was spun out of FAR.AI [107], or with the Center for AI Safety (safe.ai), which publishes benchmarks such as Humanity's Last Exam [145].

The same test applies to other comparators. The Future of Life Institute's AI Safety Index is an expert-panel rating of frontier developers, with a Summer 2026 edition covering nine companies [116]. It is consequential for reputation but is not a certification. The AI Alliance states that its Trust and Safety Evaluations Initiative is building evaluation catalogues rather than a standard, and feeds work into MLCommons [114]. Concordia AI publishes its own frontier AI risk management framework and annual State of AI Safety in China reports, which are research reports but are consultancy publications, not standards [117]. A reported "Kissinger Dialogues" working group on agentic AI security could not be located in any public source and is excluded.

4. Artifact Map: Many Layers, Not One Winner

4.1 The trust and standards stack

Agentic systems introduce a layered chain of trust that no single artifact covers. An upstream model is wrapped by an agent framework. The agent receives an identity and delegated authority from a human or organizational principal. It runs in a workload environment, calls tools and other agents through protocols, and operates under enterprise governance and external assurance, all within regulatory obligations. Figure 4.1 places the principal artifacts on that stack. The figure shows where the major artifacts sit, not their quality, and several artifacts span more than one layer.

Figure 4.1: "Many layers, not one winner" – principal AI and agentic artifacts by stack layer (status as of 2026-10-27)

LAYER	FORMAL STANDARDS (approved)	FRAMEWORKS /
CATALOGS / TAXONOMIES	PROTOCOLS / DRAFTS / TOOLS	ASSURANCE /
OBLIGATION		
-----+-----+-----		
-----+-----+-----		

7 Regulatory obligation	EN 18286:2026 (harmonised, Art.17)	GPAI Code of
Practice; MAS/HKMA guidance	prEN 18282, prEN 18228 (drafts)	EU AI Act;
China CAC measures;		
	GB 45438-2025 (CN, mandatory)	TC260 AI Safety
Gov. Framework 3.0	CN mandatory agent std (drafting)	Korea AI Basic
Act		
6 Enterprise governance &	ISO/IEC 42001:2023, 42005, 42006	CSA AICM v1.1;
AIUC-1; NIST AI RMF/600-1;	NIST COSIS overlays (pre-draft)	ISO 42001
certs (ANAB/UKAS);		
assurance	ISO/IEC 27090:2026 (guidance)	FS AI RMF; BSI
AIC4; FINOS AIGF v2		STAR for AI
L1/L2; AIUC-1 cert		
5 Agent interoperability	ITU-T F.748.93 (08/2026)	OWASP Agentic Top
10; CoSAI MCP security	A2A v1.0.0; MCP 2026-07-28; IEEE P3394	(none)
4 Agent runtime	ETSI EN 304 223 (baseline, not	OWASP ACS v0.1.0;
CSA AARM v1.0;	ETSI TS 104 328 (work item); NVIDIA	AIUC-1 quarterly
testing		
	agent-specific)	CSA Singapore
agentic addendum	OpenShell; TRACE v0.2 (dev preview)	(only
recurring agent-level		
	technical test found)	
3 Identity & delegation	OAuth 2.0 RFCs; OIDF SSF/CAEP 1.0;	CSA ATF v0.9.1
(draft); OIDF agentic	draft-ietf-wimse-aims-00; OAuth identity	(none agent-
specific)		
	AuthZEN 1.0; W3C VC 2.0; SPIFFE	identity
whitepaper	chaining (RFC Ed. queue); CoSAI ODIS; AAuth	
2 Agent builder / framework	IEEE 3119-2025 (procurement)	MAESTRO; CoSAI
agentic design patterns;	AGENTS.md; goose; Eclipse LMOS/ADL	(none)
		AWS agentic scoping
matrix; Google SAIF		
1 Upstream model	ISO/IEC 22989, 23894, 5338, 25059	MITRE ATLAS
v2026.09; NIST AI 100-2 E2025;	OpenSSF model signing v1.0; Inspect;	Frontier
safety policies (RSP,		
		AICM MDS domain
AILuminate v1.0	Preparedness, FSF); EU GPAI Code	

Three patterns stand out. First, approved formal standards are densest at the bottom and top of the stack (model lifecycle concepts and management systems) and thinnest in the middle (identity, runtime, interoperability), which is where agentic risk concentrates. Second, the middle layers are populated by fast-moving open protocols and drafts whose governance is younger than their adoption. Third, the assurance column is nearly empty for layers 3 through 5. A customer can verify that a vendor has a certified AI management system, but generally cannot verify, from a public registry, that a specific agent's delegation and runtime controls were independently tested.

4.2 Artifact feature matrix

Table 4.1 records which agent-relevant features each major artifact addresses in its published text. It is a coarse reading intended to expose gaps and overlaps, not a compliance judgment. "●" means the topic is addressed with specific, assessable requirements; "◐" means it is addressed in principle or partially; "○" means not addressed or not found.

Table 4.1: Agent-relevant feature coverage in published text (analyst reading, as of 2026-10-27)

Feature	AICM v1.1 (self-assessed)	AIUC-1 (Q4 2026)	ISO/IEC 42001	NIST AI RMF + 600- 1	ETSI EN 304 223	OWASP Agentic Top 10 + ACS	MCP / A2A specs	WIMSE AIMS / OAuth drafts	Singapore MGF Agentic
Governance and accountability	●	●	●	●	◐	○	○	○	●
Secure model lifecycle	● (MDS)	◐	◐	◐	●	○	○	○	◐
Agent identity	◐ (IAM- 12, IAM- 18)	◐	○	○	○	◐	◐ (A2A signed Agent Cards)	●	◐
Delegation chain and permission narrowing	◐ (IAM- 05, IAM- 18)	◐	○	○	○	◐	◐ (OAuth 2.1 profile)	● (draft)	●
Tool / MCP governance	◐ (AIS-11)	●	○	○	○	●	●	◐	●
Inter-agent interaction	◐	◐	○	○	○	●	●	◐	◐
Sandboxing / isolation	● (AIS-13)	◐	○	○	◐	◐	○	○	◐
Runtime enforcement and guardrails	● (TVM- 13, LOG- 15/16)	●	○	◐	◐	● (ACS)	○	○	●

Feature	AICM v1.1 (self-assessed)	AIUC-1 (Q4 2026)	ISO/IEC 42001	NIST AI RMF + 600- 1	ETSI EN 304 223	OWASP Agentic Top 10 + ACS	MCP / A2A specs	WIMSE AIMS / OAuth drafts	Singapore MGF Agentic
Kill switch / halt and recovery	◐ (BCR, SEF)	◐	○	◐	◐	◐	○	◐ (revocation)	●
Credential management	● (IAM- 02, IAM- 14, CEK)	◐	○	○	◐	◐	◐	●	◐
Evidence provenance / audit trail	● (LOG- 12, STA- 09)	◐	◐	◐	◐	◐	○	◐	●
Human oversight / escalation	● (GRC- 15)	●	◐	●	◐	◐	◐	○	●
Red teaming / adversarial testing	● (MDS- 06, TVM- 07)	● (quarterly)	○	◐	◐	◐	○	○	◐
High-risk / catastrophic actions	◐	◐	○	◐	○	◐	○	○	●
Certification pathway	● (STAR for AI)	●	●	○	○	○	○	○	○
Mandatory compliance link	◐ (mapped to EU AI Act)	◐ (mapped)	◐ (via EN 18286 context)	○	◐ (EN status)	○	○	○	○

Sources: AICM control titles from the AICM v1.1 implementation guidelines [1]; AIUC-1 [28]; ISO/IEC 42001 [36]; NIST [74][75]; ETSI [42] [43]; OWASP [20][21]; MCP and A2A [70][71]; IETF [56][57]; Singapore [128]. The AICM v1.1 control identifiers referenced include AIS-11 (Agents Security Boundaries), AIS-13 (AI Sandboxing), IAM-18 (Agent Access Restriction), GRC-15 (Human Supervision), TVM-13 (Guardrails), LOG-15 and LOG-16 (Input and Output Monitoring), and STA-09 (Service Bill of Material). The AICM column is self-assessed, because CSA authors both AICM and this paper (Section 2.5). Its ratings are derived from AICM control titles and objectives, not from audit evidence, and should be read with that limitation in mind.

On this reading, the control catalogs (AICM, AIUC-1) and the Singapore framework span the most features, although the AICM ratings carry the self-assessment caveat noted above. Protocols and identity drafts are deep on a few features and silent on governance. Management-system standards are silent on agent mechanics by design. These artifacts are therefore mostly complementary, and the main overlap is between the control catalogs themselves. We recommend that an enterprise program draw on at least one artifact from each column family.

5. Detailed Profiles

The 23 profiles below met the selection rule in Section 2.4. Each opens with a one-line profile card covering type, lanes, core artifacts with version and status, governance, verified adoption, and relationship to CSA. The narrative that follows explains significance and open questions. Organizations not profiled appear in Appendix A with the same fields in condensed form.

5.1 ISO/IEC JTC 1 (SC 42, SC 27, SC 40)

Formal SDO | L1, L3, L7, L8 | ISO/IEC 42001:2023, 42005:2025, 42006:2025, 23894:2023, 27090:2026, 27091 (FDIS) | National-body voting | Accredited certification ecosystem | Complement and foundation for CSA STAR for AI Level 2.

ISO/IEC JTC 1 Subcommittee 42 owns the most widely referenced AI management-system standard, ISO/IEC 42001:2023, which specifies requirements for establishing and continually improving an AI management system [36]. Two 2025 companions matter for assurance. ISO/IEC 42005:2025 covers AI system impact assessment [37]. ISO/IEC 42006:2025 sets requirements, on top of ISO/IEC 17021-1, for bodies that audit and certify AI management systems, and it gives accreditation bodies a common yardstick [38]. Subcommittee 27, which owns the ISO/IEC 27000 security family, produced ISO/IEC 27090, guidance on addressing security threats and compromises to AI systems. It is listed by the ANSI webstore as ISO/IEC 27090:2026 with an October 2026 edition date, while the ISO catalogue page still showed "in publication" during research [39]. It is guidance and not certifiable. The privacy companion ISO/IEC 27091 is at the FDIS approval stage [40]. ISO/IEC 38507:2022 on governance implications of AI belongs to SC 40 (IT governance), not SC 42. No public SC 42 work item specifically on AI agents was found. The nearest items are the ISO/IEC 42119 testing series, including approved work items on red teaming, and an amendment programme for the ISO/IEC 22989 terminology standard [41].

ISO/IEC's influence is likely to last, because it rests on national-body voting, regulatory referencing (the EU harmonised standards programme draws heavily on SC 42 work) and an accredited certification market. Its limitation for agentic AI is pace and abstraction. A management-system standard asks whether an organization governs AI systematically, not whether a particular agent's delegation is safe. CSA's STAR for AI Level 2 already builds on ISO/IEC 42001 certification rather than competing with it [3], which, in our assessment, makes ISO/IEC the clearest complement to CSA's assurance work.

5.2 European Union: AI Act, AI Office and CEN-CENELEC JTC 21

Government plus regional SDO | L1, L10, L11 | Regulation (EU) 2024/1689; Digital Omnibus (reported Reg. 2026/1744); GPAI Code of Practice (July 10, 2025); EN 18286:2026 | Binding law; harmonised standards confer presumption of conformity | Mandatory for in-scope systems | Mapping target for AICM.

The EU AI Act entered into force on August 1, 2024, with prohibitions applying from February 2025 and general-purpose AI obligations from August 2, 2025 [48]. The Digital Omnibus, Regulation (EU) 2026/1744, was published in the Official Journal on July 24, 2026 and applies from July 27, 2026. It delays high-risk obligations to December 2, 2027 for standalone (Annex III) systems and August 2, 2028 for systems embedded in regulated products [49]. The voluntary General-Purpose AI Code of Practice was published on July 10, 2025 and gives signatories a recognized route to demonstrate compliance with GPAI obligations [50].

CEN-CENELEC Joint Technical Committee 21 is developing the harmonised standards that will give presumption of conformity. The first, EN 18286:2026 on quality management systems for EU AI Act regulatory purposes, was published in July 2026 [46]. It was cited in the Official Journal by Commission Implementing Decision (EU) 2026/2274 on October 14, 2026, covering Article 17 obligations only [47]. The cybersecurity standard prEN 18282 completed public enquiry in mid-2026 and is in comment resolution. The risk-management standard prEN 18228 remains a draft. No harmonised standard addresses agents specifically. The EU's influence on the global landscape works through market access more than through technical novelty. Major control catalogs in this survey, including AICM v1.1 [1] and AIUC-1 [28], publish EU AI Act mappings.

5.3 NIST and the Center for AI Standards and Innovation (CAISI)

Government (U.S.) | L1, L3, L4, L7, L12 | AI RMF 1.0 (Jan 2023; under revision); AI 600-1 (July 2024); AI 100-2 E2025; AI Agent Standards Initiative (Feb 17, 2026); NCCoE agent identity concept paper (IPD Feb 2026); IR 8596 Cyber AI Profile (IPRD Dec 2025); COSIS overlays | Public comment; non-binding | Widely referenced by U.S. sector regulators and frameworks | Complement; AICM maps to AI RMF and AI 600-1.

NIST's AI Risk Management Framework remains the most cited voluntary governance framework in the United States, and NIST states that it is being revised under the White House AI Action Plan [74]. The Generative AI Profile (NIST AI 600-1) followed in July 2024 [75]. The adversarial machine learning taxonomy NIST AI 100-2 E2025 was updated in March 2025 [80]. In June 2025 the U.S. AI Safety Institute was reconstituted as CAISI [81]. On February 17, 2026, NIST announced an AI Agent Standards Initiative led by CAISI with NIST's Information Technology Laboratory and the National Science Foundation [76]. It has three pillars: support for industry-led agent standards, open-source agent protocols, and research on agent security and identity. Inputs included a CAISI request for information on AI agent security and an NCCoE concept paper on software and AI agent identity and authorization, whose comment period closed on April 2, 2026 [77]. NIST also published a preliminary draft Cyber AI Profile (NIST IR 8596) built on CSF 2.0, organized around securing AI components, AI-enabled cyber defense and thwarting AI-enabled attacks [78]. Its SP 800-53 control overlays project, renamed COSIS in October 2026, lists single-agent and multi-agent use cases but had no public draft overlay at the cutoff [79].

NIST's distinctive power is convening and referencing. Its outputs become the vocabulary of U.S. sector regulators and of frameworks such as the Treasury FS AI RMF [31]. Most agent-specific NIST work is, however, still at concept-paper or pre-draft stage. Through the industry-led pillar of its agent initiative, NIST has signaled a preference for industry-led standards [76], which leaves room for open foundations and consortia to supply the technical content.

5.4 ETSI Technical Committee Securing AI (TC SAI)

Formal SDO (European, global membership) | L2, L4, L6 | TS 104 223 (Apr 2025) → EN 304 223 V2.1.1 (Dec 2025); TR 104 128; TS 104 275 (Sept 2026); TS 104 328 agentic work item (adopted Oct 1, 2026) | ETSI member process | Basis of UK government code | Complement; potential reference baseline for agent controls.

ETSI TC SAI converted the UK government's January 2025 Code of Practice for the Cyber Security of AI into ETSI TS 104 223. That specification sets out 13 principles expanded into 72 provisions across five lifecycle phases [42][45]. In December 2025 it was elevated to European Standard EN 304 223 V2.1.1 [43]. On October 1, 2026, TC SAI adopted work item DTS/SAI-001 for TS 104 328, "Baseline Cyber Security Requirements for Agentic AI Systems," which extends EN 304 223 across the agentic lifecycle [44] **[VERIFICATION NEEDED: no public ETSI work-programme record located]**. The work item is supported by the UK NCSC and DSIT, among others. It is a work item with no published draft and no target date, and should not be described as a standard. ETSI is significant because it is the only *European* formal SDO with an adopted work item for baseline agentic AI cybersecurity requirements, and because EN status gives it a path into European procurement and regulation. Comparable efforts are under way in China (TC260), at ITU-T and at IEEE, at varying stages (Sections 5.5, 5.17 and 5.19).

5.5 ITU-T (Study Groups 17, 20 and 21)

Intergovernmental SDO | L4, L5 | Rec. ITU-T F.748.93 (08/2026) on AI agent interoperability; Focus Group on Trust and Identity for Humans and Agentic AI (launched July 9, 2026); X.sg-AIAof (under study, target June 2028); Y.El-AIS (under study) | Member-state and sector-member process | Early | Monitor and contribute.

ITU-T Study Group 21 approved Recommendation F.748.93, "Framework and Requirements for AI Agent Interoperability," on August 29, 2026 [52]. This makes it one of the few approved formal standards with "agent" in its scope. On July 9, 2026, ITU launched a Focus Group on Trust and Identity for Humans and Agentic AI under Study Group 17 [53]. Its mandate covers terminology, reference architectures for agent identity and discovery, trust frameworks, lifecycle assurance models and a standardization roadmap, with a first meeting in Paris in November 2026. SG17 is also studying X.sg-AIAof, security guidelines for cloud-side AI agent orchestration frameworks, registered in June 2026 with a target of June 2028 [54]. SG20 is studying Y.El-AIS, an evaluation framework for IoT AI agent services [55]. Editors on several of these items are Chinese contributors, and CAICT is listed as a supporter on related drafts **[VERIFICATION NEEDED: cite ITU-**

T work-programme records directly] ITU-T is therefore a principal venue through which Chinese agent-standardization work enters the international system. Focus groups produce technical reports rather than Recommendations, so the Focus Group's influence will depend on uptake by SG17.

5.6 IETF (OAuth and WIMSE working groups)

Formal SDO (open participation) | L4, L5 | draft-ietf-wimse-aims-00 (Sept 15, 2026); draft-ietf-oauth-identity-chaining (approved, RFC Editor queue); draft-ietf-oauth-transaction-tokens-11; OAuth 2.1 draft; individual drafts including AAuth (-11) | Rough consensus | OAuth RFCs are ubiquitous; agent drafts unproven | Complement; potential home for CSA delegation-evidence ideas.

The IETF is where agent authorization is most likely to be standardized at internet scale, but at the cutoff **no agent-specific IETF document was an RFC** [56][57]. The WIMSE working group adopted draft-ietf-wimse-aims-00, "AI Identity Management System," on September 15, 2026 [56]. Its authors include engineers from AWS, OpenAI and Okta, and it applies workload identity and OAuth mechanisms to agents rather than defining a new protocol. In the OAuth working group, the identity-chaining draft (cross-domain token exchange) has been approved and is in the RFC Editor queue. The transaction-tokens draft has reached working-group consensus, and OAuth 2.1 itself remains a draft [57]. Numerous individual drafts address agent delegation, including Dick Hardt's AAuth protocol (draft-hardt-oauth-aauth-protocol-11, September 25, 2026), which is an individual submission not adopted by any working group [58]. MCP's authorization model already references the OAuth 2.1 draft and several OAuth RFCs [71], which illustrates how IETF work becomes de facto infrastructure before it is finished.

5.7 OpenID Foundation, FIDO Alliance and W3C (identity cluster)

Open consortia and SDO | L4 | OIDF "Identity Management for Agentic AI" (Oct 2025); Shared Signals Framework, CAEP and RISC 1.0 final (Sept 2025); AuthZEN Authorization API 1.0 final (Jan 2026); FIDO Agentic Authentication Technical Working Group (Apr 28, 2026); W3C VC Data Model 2.0 Recommendation (May 15, 2025) | Member processes | OIDC widely deployed; agent work early | Complement.

The OpenID Foundation's AI Identity Management Community Group published a whitepaper on identity management for agentic AI in October 2025 [64]. The Foundation finalized the Shared Signals Framework with CAEP and RISC in September 2025 [65] and the AuthZEN Authorization API 1.0 in January 2026 [66]. These provide revocation signalling and externalized authorization primitives that agent architectures can reuse. The FIDO Alliance launched an Agentic Authentication Technical Working Group and agent-commerce specification work on April 28, 2026, with contributions including Google's AP2 and Mastercard's Verifiable Intent [67]. No final FIDO agent specifications exist yet. W3C's Verifiable Credentials Data Model 2.0 became a Recommendation in May 2025, and DID 1.1 is a Candidate Recommendation [62]. W3C's agent-related community groups do not produce standards [63]. Together these bodies supply the building blocks for human-principal binding, revocation and high-value transaction authorization. None yet supplies an end-to-end agent delegation profile.

5.8 Linux Foundation Agentic AI Foundation (MCP, A2A, AGENTS.md)

Open foundation | L5, L4 | MCP specification 2026-07-28; A2A v1.0.0 (Mar 2026); AGENTS.md; goose; agentgateway | Linux Foundation directed fund; member governance | Very broad implementation (de facto standards) | Complement; protocol layer CSA does not and should not own.

The Agentic AI Foundation (AAIF) was announced on December 9, 2025 with three founding projects: Anthropic's Model Context Protocol, Block's goose and OpenAI's AGENTS.md [72]. On May 18, 2026 it added 43 members for a total of 190 [68], and A2A, a Linux Foundation project since June 2025, joined AAIF on August 17, 2026 [69]. A2A reached its first stable release, v1.0.0, in March 2026, with signed Agent Cards and declared security schemes [70]. The current MCP specification version is dated 2026-07-28. Its authorization section makes authorization optional, but when it is used over HTTP, servers act as OAuth 2.1 resource servers, must publish protected-resource metadata, and clients must use resource indicators [71]. By membership (190 as of May 2026) and breadth of protocol implementation, AAIF is arguably the most consequential host for agent interoperability, because it holds the protocols most agents

actually speak. Its specifications are developed through open-source project governance rather than formal consensus balloting, and they define *mechanisms*, not *assurance*. They say how an agent authenticates to a tool, not whether the deployment is adequately controlled. This is the clearest example of a layer where CSA's role is mapping and evidence rather than authorship.

5.9 OWASP GenAI Security Project

Open foundation | L3, L6, L7 | Top 10 for LLM Applications 2026 (announced Sept 2026); Top 10 for Agentic Applications for 2026 (Dec 9, 2025); Agent Control Standard v0.1.0 (Sept 2026); AIVSS v0.8; AI Exchange; MCP Top 10 (beta) | Open community; vendor sponsors | Very high practitioner reach | Overlap at taxonomy layer; strong mapping partner.

Measured by community size, OWASP's GenAI Security Project is among the most influential practitioner taxonomy publishers. In September 2026 it announced the 2026 edition of the Top 10 for LLM Applications, together with the donated Agent Control Standard (ACS), and reported that its community exceeded 30,000 members [19]. The project's own LLM Top 10 landing page still showed the 2025 list during research. The Top 10 for Agentic Applications, labeled for 2026, was published on December 9, 2025 with risks ASI01 through ASI10 [20]. ACS v0.1.0 defines runtime hooks through which a "guardian" policy authority inspects an "observed agent." Its README lists known gaps, including fail-open defaults, and a v0.2.0 is targeted for March 2027 [21]. The AI Vulnerability Scoring System (AIVSS) extends CVSS concepts to agentic AI and was at v0.8 in March 2026 [22]. The OWASP AI Exchange, a flagship project since March 2025, has a liaison with CEN-CENELEC and contributed text to ISO/IEC 27090 and prEN 18282 [23]. OWASP's taxonomies are not auditable controls, but OWASP is moving toward runtime specification (ACS) and scoring (AIVSS), which brings it closer to CSA's control and evidence layer.

5.10 Coalition for Secure AI (CoSAI), an OASIS Open Project

Open consortium | L2, L3, L4, L9, L12 | Signing ML Artifacts (Sept 2025); AI Incident Response Framework (Oct 2025); MCP Security (Jan 2026); Future of Agentic Security (Mar 2026); Agentic IAM (Apr 2026); AI Shared Responsibility Framework (May 2026); CoSAI Risk Map; ODIS (work item) | OASIS open-project governance; major vendor sponsors | Strong sponsor base; implementation evidence limited | Partial overlap (risk map, shared responsibility); natural partner.

CoSAI was founded in July 2024 as an OASIS Open Project with sponsors including Amazon, Anthropic, Cisco, Google, IBM, Intel, Microsoft, NVIDIA, OpenAI, PayPal and Wiz [24]. It runs four workstreams: AI supply chain (WS1), preparing defenders (WS2), AI security risk governance (WS3) and secure design patterns for agentic systems (WS4). Its second-year review lists papers on signing ML artifacts, incident response, MCP security, agentic identity and access management and an AI shared responsibility framework [25]. In September 2025 Google donated SAIF data to form the CoSAI Risk Map [26]. The Open Delegation and Identity Standard (ODIS) repository contains RFC-style documents and a candidate reference implementation extending OAuth 2.0, OIDC and SPIFFE for agent delegation, but states no version or release [27]. It is best treated as a work item. A "CoSAI Oracle" artifact named in the seed inventory could not be found. CoSAI's AI Shared Responsibility Framework overlaps conceptually with AICM's five-actor Shared Security Responsibility Model. Whether the two are reconcilable is an open mapping question rather than a settled conflict.

5.11 MITRE: ATLAS, AI incident sharing, and the CWE/CVE ecosystem

Not-for-profit operator of public knowledge bases | L3, L9 | ATLAS data v2026.09 (Sept 15, 2026); AI Incident Sharing (Oct 2024); CWE-1427 (prompt injection) and CWE-1426 | MITRE-governed with community input | Very broad use in threat modeling | Complement; AICM threat mapping and MAESTRO use ATLAS concepts.

MITRE ATLAS now releases monthly. Version 2026.09 lists 16 tactics, 120 techniques, 88 sub-techniques, 40 mitigations and 73 case studies [82]. Since May 2026 every technique carries a platform tag, including "Agentic AI," and recent releases added agent-specific techniques such as AI agent tool poisoning [82]. ATLAS is distributed in STIX alongside ATT&CK, and it is the only threat taxonomy reviewed here that is distributed in STIX, which makes it the easiest of them for machines to read. MITRE also launched an AI Incident Sharing initiative in October 2024 [83]. Within the CWE catalogue, CWE-1427 covers improper neutralization of input used for LLM

prompting [84]. The CVE Program, which MITRE helps operate but does not solely govern, has an AI working group and a pilot for frontier AI researcher CNAs [85]. ATLAS is the de facto adversary vocabulary for AI, but it is not a control catalog, and its mitigations are not auditable requirements.

5.12 Cloud Security Alliance and CSAI Foundation (self-assessed)

Open nonprofit | L1, L2, L3, L4, L6, L8 | AICM v1.1 (June 22, 2026; 247 controls, 18 domains); AI-CAIQ v1.1; STAR for AI Levels 1-2; Valid-AI-ted; MAESTRO; ATF v0.9.1 (public review draft); AARM v1.0; SAGE v1.0 RC1; Catastrophic Risk Annex (work item); CVE Numbering Authority | Working-group consensus; member-funded; CSAI Foundation a 501(c)(3) | STAR Registry public | Subject of this paper – conflict of interest.

CSA released AI Controls Matrix v1.1 on June 22, 2026 (announced by blog on July 14, 2026) [1][2]. It contains 247 control objectives in 18 domains, up from 243 in v1.0. The Model Security (MDS) domain of 13 controls, present since v1.0, was retained and refined. The matrix is synchronized with CCM v4.1 and governed by the Security Controls Catalog subgroup of the Compliance Automation Revolution working group. CSA's introductory guidance gives "June 2026" and "July 2026" in different passages. This paper uses the June 22 artifact date. AICM v1.1 publishes control-level mappings to ISO/IEC 42001:2023, NIST AI RMF and AI 600-1, the EU AI Act, BSI AIC4, and (new in v1.1) AIUC-1. The accompanying CSA blog reports "No Gap" coverage of 85% for BSI AIC4, 59% for ISO/IEC 42001, 28% for AIUC-1, 20% for the EU AI Act and 7% for NIST [2]. These figures measure how much of each target framework is fully covered by AICM controls, not AICM's quality. The relatively low AIUC-1 figure shows that the two catalogs differ substantially in scope. It may reflect agent-specific testing requirements in AIUC-1 that AICM does not address, breadth in AICM that AIUC-1 does not cover, or both. The low "No Gap" figures for NIST (7%) and the EU AI Act (20%) likewise show that AICM alone does not establish conformity with those frameworks. AICM assigns responsibilities across five actor roles (model providers, orchestrated service providers, application providers, cloud service providers and AI customers) and is accompanied by the 320-question AI-CAIQ.

Assurance runs through STAR for AI. Level 1 is a self-assessment AI-CAIQ, optionally scored by the AI-based Valid-AI-ted service. Level 2 combines a Valid-AI-ted AI-CAIQ with ISO/IEC 42001 certification, or with certification or attestation routes delivered by accredited bodies and CPAs [3][4]. Valid-AI-ted badges expire after 12 months. On June 30, 2026, CSA added AIUC-1 certification to the STAR Registry [5]. The CSAI Foundation, launched in March 2026 with a mission of "securing the agentic control plane," hosts agentic work [6] [7]. That work includes the Agentic Trust Framework, a v0.9.1 public review draft authored by Josh Woodruff with AICM as its "parent umbrella" [9][10]. It also includes AARM v1.0, a runtime action-management specification donated by Vanta and now a CSA working group [12][13], and SAGE, a machine-readable format for security guidance at v1.0 RC1 [14]. The Catastrophic Risk Annex for STAR for AI is a phased work item running to December 2027 [8]. CSAI also announced that it has become a CVE Numbering Authority for AI-specific vulnerabilities [7]. CSA has also published guidance on non-human identity management for agentic AI [149] and on translating zero trust frameworks into enterprise practice for agentic AI [150]. Both are relevant to the identity and delegation gap identified in this paper.

The same neutrality question this paper raises for AIUC-1 (Section 5.13) applies to CSA. CSA both authors AICM and operates the STAR for AI registry and the AI-based Valid-AI-ted scoring service, and it is funded by members, some of whom are vendors assessed through those programs. Readers should weigh CSA's position with that in mind. Finally, two artifacts are sometimes mistaken for CSA work but are not CSA's. The TRACE agent-attestation project under agenttrust-io is hosted by the Linux Foundation and has no CSA connection [73], and AAuth is an individual IETF draft [58].

5.13 AIUC-1 (Artificial Intelligence Underwriting Company)

Commercial standard owner with consortium input | L2, L7, L8 | AIUC-1, quarterly releases (latest Q4 2026, Oct 15, 2026); certification with accredited auditors | Company-owned; technical contributors include Stanford, MIT, MITRE, CSA, Google Cloud, Cisco | Public list of certified companies | Direct overlap at agent control and certification layer; also mapped and registry partner.

AIUC-1 positions itself as a standard for AI agent security, safety and reliability. It is updated quarterly; the Q4 2026 release (October 15, 2026) added open-weight model controls and stronger third-party testing [28]. Certification requires quarterly technical testing (red teaming) and an annual controls review. This recurring, agent-level testing is the feature that most clearly distinguishes it from organization-level schemes. Schellman became the first accredited AIUC-1 auditor on February 3, 2026 [29], and Coalfire holds provisional accreditation [30]. AIUC-1's site lists certified organizations including Cursor, ElevenLabs, Harvey, UiPath and KPMG [28]. It

publishes crosswalks to MITRE ATLAS, ISO/IEC 42001, the EU AI Act, NIST AI RMF and OWASP, but **not** to CSA AICM. The mapping runs in the other direction, from AICM v1.1 to AIUC-1 [1][28]. The governance model is the principal difference from consensus bodies. The standard and the certification scheme are owned by a company whose business includes insuring AI agents. That gives commercial focus and speed, but raises neutrality questions that the consortium model only partly answers. Reports of a combined SOC 2 and AIUC-1 engagement with the AICPA could not be confirmed from AICPA primary sources and are treated as unverified.

5.14 Financial-sector baselines: U.S. Treasury and CRI, MAS, HKMA, FINOS

Government and sector consortia | L2, L11 | FS AI RMF (Feb 19, 2026; 230 control objectives); MAS Guidelines on AI Risk Management (Oct 7, 2026); HKMA revised guidance (Oct 15, 2026); FINOS AI Governance Framework v2.0 (Nov 2025) | Regulator-issued or consortium | Sector-mandated or strongly expected | Overlap with AICM for financial buyers; mapping opportunity.

Among the sectors reviewed, financial services is, in our assessment, producing the most consequential AI control baselines, because several are issued or expected by regulators. The U.S. Treasury released the Financial Services AI Risk Management Framework on February 19, 2026, developed through a public-private AI Executive Oversight Group with the Cyber Risk Institute closely involved [31]. Secondary analyses describe 230 control objectives organized by NIST AI RMF functions and scaled by adoption stage **[VERIFICATION NEEDED: CRI primary page not retrieved]**. The Monetary Authority of Singapore issued final Guidelines on AI Risk Management on October 7, 2026, effective October 2027, and plans to consult on agentic AI guidance in 2027 [99]. The Hong Kong Monetary Authority issued revised guidance under an "ART" (Accountability, Resilience, Trust) framework on October 15, 2026 [100]. FINOS, a Linux Foundation project, released v2.0 of its AI Governance Framework in November 2025 with an agentic risk catalogue [98]. AIUC-1 announced a collaboration with CRI in August 2026, but its details are not public [28]. For CSA, these baselines are the most likely substitutes in financial-sector procurement, and also the most valuable mapping targets.

5.15 Singapore: IMDA, AI Verify Foundation and the Cyber Security Agency

Government and open foundation | L1, L2, L6, L7, L11 | Model AI Governance Framework for Agentic AI (Jan 22, 2026); Guidelines on Securing AI Systems with agentic addendum (final June 17, 2026); AI Verify toolkit; Global AI Assurance Pilot; Starter Kit for LLM testing | Government-led with international partners | Active pilots | Complement; strong partnership candidate.

Of the government frameworks reviewed, Singapore's covers the most agent-specific features in Table 4.1. IMDA launched a Model AI Governance Framework for Agentic AI at Davos on January 22, 2026 [128]. It is organized around four dimensions: bounding risks upfront by limiting agents' tools and scope, keeping humans meaningfully accountable, applying technical controls across the lifecycle, and enabling end-user responsibility. The Cyber Security Agency of Singapore finalized a "Securing Agentic AI" addendum to its AI security guidelines on June 17, 2026 [129]. The AI Verify Foundation runs open-source testing tooling and ran a Global AI Assurance Pilot in 2025 that paired specialist testers with deployers across industries [130]. Singapore also convened the Singapore Consensus on Global AI Safety Research Priorities in 2025 [131]. Singapore's frameworks are voluntary but specific about controls, such as limits on agents' tools and scope. Their weakness as a global baseline is that they are not certifiable and depend on other schemes for assurance, which makes them a natural complement to STAR for AI and AICM.

5.16 United Kingdom: DSIT, NCSC and the AI Security Institute

Government | L2, L7, L10, L12 | Code of Practice for the Cyber Security of AI (Jan 31, 2025), now ETSI EN 304 223; AISI Inspect; Inspect Cyber (June 2025); International AI Safety Report secretariat | Government with open-source tooling | Inspect widely used | Complement.

The UK's main standard-setting contribution was to route its Code of Practice for the Cyber Security of AI into ETSI, where it became TS 104 223 and then EN 304 223 [42][45]. The UK AI Security Institute (renamed from the AI Safety Institute in February 2025) maintains Inspect, an open-source evaluation framework, and Inspect Cyber for agentic cyber evaluations [126][127]. AISI also provides the secretariat for the International AI Safety Report, whose 2026 edition was published in February 2026 [136]. The UK model is to use formal SDOs for requirements and government institutes for evaluation tooling, a separation that the commissioning brief's layered thesis anticipates.

5.17 China: TC260, CAC and CAICT

National SDO, regulator and state research institute | L1, L2, L4, L5, L11 | TC260 AI Safety Governance Framework 3.0 (Sept 2026); GB 45438-2025 (mandatory content labelling); GB/T 45654-2025 (generative AI service security); agent deployment security guidance (July 2026); mandatory national standard project on AI agent application security (listed June 27, 2026) | State-coordinated | Binding for in-scope services in China | Monitor; limited direct overlap but high relevance to global convergence.

China's National Cybersecurity Standardization Technical Committee (TC260), working under the Cyberspace Administration of China, published AI Safety Governance Framework 2.0 in September 2025. It released Framework 3.0 in September 2026, shifting emphasis toward AI agents, embodied systems and loss of human control [148] (secondary sources disagree on whether the date was September 14 or 24). Framework documents are non-binding, but they are translated into national standards. GB 45438-2025 on labelling AI-generated content is mandatory and took effect on September 1, 2025. GB/T 45654-2025, a recommended standard on basic security requirements for generative AI services, took effect on November 1, 2025 [138]. Most relevant to agents, a mandatory national standard project on general security requirements for AI agent applications was listed in June 2026 on an 18-month cycle. Its planned requirements cover agent identity, permissions, tool use, human intervention for high-risk operations, logging and emergency shutdown [139] **[VERIFICATION NEEDED: reported by CGTN; TC260 primary text not retrieved]**. CAICT, a research institute under the Ministry of Industry and Information Technology, supports agent evaluation drafts in ITU-T. If the reported mandatory standard project [139] is confirmed and completed on its 18-month cycle, China would be the first major jurisdiction with a *mandatory*, agent-specific runtime security standard, likely ahead of any comparable Western requirement.

5.18 Japan, Korea, India and Australia

Government | L1, L7, L11 | Japan AI Promotion Act (May 2025) and AISI Evaluation Perspectives Guide v1.20 (July 2026, adds AI agents); Korea AI Basic Act (in force Jan 22, 2026); India AI Governance Guidelines (Nov 5, 2025); Australia Guidance for AI Adoption (Oct 2025) and Australian AISI | Government | Varies | Complement.

Across the rest of Asia-Pacific, governments are building institutes and guidance rather than private consortia. Japan's AI Safety Institute updated its Guide to Evaluation Perspectives on AI Safety to v1.20 in July 2026, adding AI agents [140]. Korea's AI Basic Act took effect on January 22, 2026, with an enforcement decree, domestic-representative obligations for large foreign providers, and a grace period of at least one year before most fines [146]. India released voluntary AI Governance Guidelines on November 5, 2025, built on seven principles and proposing new coordinating bodies [147]. The Bureau of Indian Standards has adopted ISO/IEC 42001 nationally. Australia replaced its 2024 Voluntary AI Safety Standard with Guidance for AI Adoption in October 2025 and announced an AI Safety Institute in November 2025 [142]. These jurisdictions generally adopt ISO/IEC standards and contribute evaluation capability through the International Network for Advanced AI Measurement, Evaluation and Science (formerly the International Network of AI Safety Institutes) [142]. They do not issue rival control catalogs.

5.19 IEEE Standards Association

Formal SDO | L1, L5, L8 | IEEE 3119-2025 (AI procurement); P3394 LLM Agent Interface (in sponsor ballot); P3709, P3777, P3931 (agent PARs); IEEE 7001-2021; CertifAIEd | IEEE-SA process | Moderate | Monitor.

IEEE published IEEE 3119-2025 on procurement of AI and automated decision systems in May 2025 [61]. It has at least four agent-related standards projects under way. P3394, a standard for an LLM agent interface, has reached the ballot invitation stage [60]. Further project authorization requests cover agentic AI frameworks (P3709), agent benchmarking (P3777) and agent description, discovery and registry interoperability (P3931) [59]. None of these is published. IEEE also operates CertifAIEd, an ethics-oriented certification programme. IEEE's agent projects overlap with AAIIF protocols and ITU-T F.748.93. Which of these becomes authoritative for agent interfaces is one of the clearest fragmentation risks in the landscape (Section 9).

5.20 MLCommons and independent evaluators (METR, Apollo Research, FAR.AI)

Benchmark consortium and research nonprofits | L7, L10 | Alluminate v1.0 (Dec 2024) and v1.1 (French); jailbreak benchmark v0.5 (Oct 2025); Agentic Product Maturity Ladder v0.1 (Feb 2026); METR time-horizon and frontier risk reports; Apollo scheming evaluations | Consortium and nonprofit | Used by labs and governments | Complement; evidence suppliers.

MLCommons released the Alluminate v1.0 safety benchmark for general-purpose chat models in December 2024 [103]. Its agentic workstream had produced only a release-candidate Agentic Product Maturity Ladder (v0.1) by February 2026 [104]. MLCommons calls its benchmarks standards, but it is a benchmark consortium rather than a formal SDO. METR publishes methodology for measuring how long agents can work autonomously, and in 2026 piloted a Frontier Risk Report with participation from major developers [105]. Apollo Research specializes in evaluations of strategic deception and scheming, including work with OpenAI [106]. FAR.AI leads a consortium building CBRN evaluations for the EU AI Office [107]. These organizations generate evidence that assurance schemes could consume. None operates a certification scheme, and none should be classed as a standards competitor.

5.21 Frontier Model Forum and frontier developer frameworks

Industry body and corporate policies | L10, L12 | FMF issue briefs including "Emerging Security Practices for AI Agents" (June 2026) and "Agents for Cyber Defense" (Sept 2026); information-sharing agreement (Mar 2025); Anthropic RSP v3.4 (July 8, 2026); OpenAI Preparedness Framework v2 (Apr 2025); Google DeepMind Frontier Safety Framework v3.1 (Apr 2026) | Member-only industry body; company-owned policies | Applied by signatories | Complement; CSA Catastrophic Risk Annex intersects.

The Frontier Model Forum publishes technical reports and issue briefs on frontier safety frameworks, capability thresholds and, in 2026, security practices for AI agents and agents for cyber defense [108]. Its members signed an information-sharing agreement covering threats, vulnerabilities and capability advances in March 2025 [109]. The frontier developers' own frameworks are versioned corporate policies. Anthropic's Responsible Scaling Policy is at v3.4, effective July 8, 2026 [110]. OpenAI's Preparedness Framework v2 dates from April 2025 [111], and Google DeepMind's Frontier Safety Framework is at v3.1 (April 2026) [112]. These policies govern model release decisions and are increasingly referenced by regulation, including the EU GPAI Code of Practice. They do not address enterprise agent deployment controls. CSA's Catastrophic Risk Annex, if completed, would occupy the space between frontier developer policies and enterprise assurance.

5.22 Vendor and vendor-originated frameworks

Commercial vendors | L2, L3, L5, L6 | Google SAIF with agent risk map; Microsoft Taxonomy of Failure Modes in Agentic AI Systems v2.0 (June 2026); NVIDIA OpenShell (Mar 2026) and Open Agent Safety Platform (Sept 2026); AWS Agentic AI Security Scoping Matrix (Nov 2025); Cisco Integrated AI Security and Safety Framework (Dec 2025); Vanta Agentic Trust Controls; Pillar Security SAIL 2.0 (July 2026) | Vendor-owned, several open-licensed | Broad within each vendor's customer base | Mixed: inputs to neutral bodies, and in two cases (ATC, SAIL) competing control sets.

Large vendors publish influential frameworks that often seed neutral bodies. Google's SAIF data became the CoSAI Risk Map [26][92]. Microsoft's AI Red Team updated its taxonomy of agentic failure modes to v2.0 in June 2026, adding modes such as inter-agent trust escalation and MCP or plugin abuse [94]. NVIDIA open-sourced OpenShell, an agent runtime with a sandbox, policy engine and privacy router, in March 2026 [95]. In September 2026 it announced an Open Agent Safety Platform that pairs OpenShell with a hardware watchdog able to quarantine agents [96]. AWS's Agentic AI Security Scoping Matrix classifies agents by four levels of agency [97], and Cisco published an integrated AI security and safety taxonomy [93]. Two vendor-originated catalogs compete more directly with neutral control catalogs. Vanta's Agentic Trust Controls (65 controls, reportedly mapped to AICM and AIUC-1 among others) [32] **[VERIFICATION NEEDED: available sources confirm ISO/IEC 27001 and 42001 mappings only; the 12-domain count belonged to an earlier 61-control version]** and Pillar Security's SAIL 2.0 (91 risks across seven lifecycle phases) [33] both target the same buyers as AICM and AIUC-1. If the AICM mapping is confirmed, it would indicate that these vendors position their catalogs as complements. Even so, buyers face a proliferation of catalogs. Vanta's relationship with CSA runs both ways: Vanta donated AARM to CSA (Section 5.12), while its ATC is a competing catalog.

5.23 Software supply chain: OpenSSF, SLSA, Akrites, CIS and SAFECode

Open foundations | L9, L12 | OpenSSF Security-Focused Guide for AI Code Assistant Instructions (Aug 2025); OpenSSF Model Signing v1.0 (Apr 2025); SLSA v1.2; Akrites (launched June 25, 2026); CIS/SAFECode Secure by Design v1.1 (July 2026) | Linux Foundation and nonprofit governance | Broad open-source use | Complement.

AI-generated code and model supply chains are addressed mainly by open-source security foundations. OpenSSF published guidance on security-focused instructions for AI code assistants [87] and, with Sigstore, released v1.0 of a model-signing library [88]. SLSA v1.2 is the current supply-chain levels specification, though it contains no AI-specific guidance [89]. On June 25, 2026, the Linux Foundation launched Akrites, a shared incident-response and coordinated-disclosure effort for critical open source facing AI-enabled threats, with founders including major AI developers and banks [90]. CIS and SAFECode released Secure by Design v1.1 with expanded AI content in July 2026 [91]. No current affiliation between SAFECode and CSA was found beyond historical collaboration. This lane remains thinly specified for agent-authored code, a gap discussed in Section 10.

Editorial Status: Publication Hold

Do not publish or generate a PDF from this draft. The source manuscript ended after Section 5.23. Sections 6 through 11 are missing, including the assurance schemes table (Table 8.1), the board and CISO checklist, the standards-power assessment (Section 9.4), the gap analysis (Section 10) and the CSA strategy chapter with its counterarguments (Section 11). Appendices A through C and the full reference list are also missing. As a result, the census counts, the top-ten ranking and the gap hypotheses in the Executive Summary cannot yet be checked against their supporting material. The style edit and the quality-review corrections have been applied to the surviving text. Restore the missing material from the original drafting run, re-run citation validation, and remove this section before publication.

References

This reference list is incomplete. It contains only the entries that the citation review confirmed against public sources, numbered to match the citations in the text. Reference numbers [1]–[148] that are not listed below must be restored from the original manuscript. Each restored entry needs a public URL, and no entry may point to an internal or unpublished source.

- [1] Cloud Security Alliance. "[AI Controls Matrix \(AICM\) v1.1](#)." Cloud Security Alliance, June 2026.
- [2] Cloud Security Alliance. "[AI Controls Matrix v1.1: Strengthening the Foundation for Trustworthy AI](#)." CSA Blog, July 14, 2026.
- [5] Cloud Security Alliance. "[CSA Extends Leadership into Agentic AI with Addition of AIUC-1 Certification to STAR Registry](#)." Press release, June 30, 2026.
- [19] OWASP GenAI Security Project. "[OWASP GenAI Security Project Unveils 2026 Top 10 for LLM Applications, New Agent Control Standard and Sponsors as Community Tops 30,000 Members](#)." OWASP, September 1, 2026.
- [32] RuntimeWire. "[Vanta Agentic Trust Controls for AI Agent Governance](#)." RuntimeWire, 2026. (Secondary source; primary source to be confirmed.)
- [47] Digital Policy Alert. "[Harmonised Standard on Quality Management System for EU AI Act Regulatory Purposes Entered into Force \(EN 18286:2026\)](#)." Digital Policy Alert, October 2026.
- [49] National Law Review. "[EU Digital Omnibus on AI Enters Force](#)." The National Law Review, July 2026.
- [52] International Telecommunication Union. "[Recommendation ITU-T F.748.93: Framework and Requirements for AI Agent Interoperability](#)." ITU-T, August 2026.
- [56] IETF WIMSE Working Group. "[AI Identity Management System \(draft-ietf-wimse-aims\)](#)." Internet-Draft, September 15, 2026.
- [68] Linux Foundation. "[Agentic AI Foundation Adds 43 New Members as Enterprise and Government Adoption of Open Agent Standards Accelerates](#)." Press release, May 18, 2026.
- [69] Agentic AI Foundation. "[A2A Joins AAIF](#)." AAIF Blog, August 17, 2026.
- [76] National Institute of Standards and Technology. "[AI Agent Standards Initiative](#)." NIST CAISI, February 17, 2026.
- [79] National Institute of Standards and Technology. "[Control Overlays for Super Intelligence Systems \(COSIS\)](#)." NIST Computer Security Resource Center, October 20, 2026.
- [149] Cloud Security Alliance. "[Non-Human Identity Management for Agentic AI: Extending IAM Beyond Humans](#)." Cloud Security Alliance.
- [150] Cloud Security Alliance. "[Zero Trust for Agentic AI: Translating Frameworks into Enterprise Practice](#)." Cloud Security Alliance.