

What the 2026 AI Governance Frameworks Bind, and What They Leave Open

Pharns Genece

AAM Cyber · ORCID 0009-0009-0791-4124

A survey of the assessment and enforcement literature

October 6, 2026

Technical Report AAM-TR-2026-01

Abstract

Two bodies of work on AI governance developed through 2025 and 2026, and they are often read as one conversation. The first concerns assessment: who evaluates a system, under what conditions, and how the result is reported. The second concerns enforcement: what a system is permitted to do at the moment it acts. This survey describes what each body of work specifies, taken from the documents' own scope statements rather than inferred, and identifies where the two stop. The assessment literature produces conditions under which an evaluation may be called independent. The enforcement literature produces controls that bind an action before it occurs. Outside one recently proposed framework scoped to financial services, neither produces a result from an examination that a reader can set beside another result and compare across organizations or sectors. Otherwise the two literatures approach the same territory from opposite sides and stop at the same boundary, from opposite directions.

Keywords: AI governance, third-party assessment, evaluator independence, runtime enforcement, conformity assessment, AI assurance

1. Scope of this survey

This is a description of published work, not a proposal. It does not argue that any particular institution should be built, and it does not specify what a measurement of governance quality would have to satisfy. It reports what the documents say and what their authors state they were scoped to do.

Where this survey identifies something a body of work does not address, that observation is taken from the work's own scope section rather than inferred from its absence. The distinction matters: a framework written to specify operating conditions has not omitted accreditation, it has scoped itself to something else.

2. The assessment literature

In December 2025 a group of evaluation organizations published **AEF-1, Minimum Operating Conditions for Independent Third Party AI Evaluations** [1]. It sets out five conditions, each with specific requirements and recommendations: sufficient access and resources, minimized conflicts of interest, analytic autonomy, transparent methods and results, and protection of sensitive information. It is accompanied by a checklist an evaluator completes and publishes alongside its results.

The document is more operational than most discussion of this subject. It prohibits compensation contingent on findings. It bars the system provider from owning voting shares or a large proportion of shares in the evaluator, and from controlling any of the evaluator’s board seats. It requires the evaluator to determine its own metrics, elicitation strategy, and scoring rubrics. It requires the evaluator to retain editorial control over how results are presented, and it removes the provider’s authority to redact results in order to conceal concerning findings. It requires disclosure of funding relationships and recusal of individuals with financial interests in the evaluated party. Its conflict-of-interest categories, in its own account, are “broadly similar to those in regulations on credit rating agencies and financial auditing,” and it cites the auditing profession’s independence rule as an example [11].

Most of what an abstract account of evaluator independence would propose is therefore already specified, in more detail, by people who evaluate frontier systems for a living. A reader who wants the conditions should read AEF-1.

A second contribution appeared in January 2026. A multi-author report on frontier AI auditing [2], with authors drawn from AI evaluation and research organizations, universities, and policy institutes, proposed an auditor accreditation program with tiered certifications and an auditor-of-auditors body with authority to revoke accreditation. Its treatment of auditor independence specifies mandatory disclosure of financial relationships, standardized engagement terms designed to prevent auditor-shopping, cooling-off periods in both directions, and alternative payment models that reduce auditor dependence on auditees. An appendix is devoted to the risks of auditee-pays arrangements.

A third contribution, in October 2026, goes further than either on the question this paper is asking, and it is worth stating plainly rather than filing under adjacent work. A vertical framework for financial services [4] proposes a per-agent unit of assurance evidence recording a bounded, falsifiable claim about whether a system operates within a defined boundary. Assurance progresses through three levels against a constant schema: a profile asserted from institutional evidence, a profile validated by testing the deployment in its own environment, and a profile verified by a qualified independent assessor running the same tests. Alongside these it defines a benchmarked profile, which in the authors’ words “reports results comparable across institutions under reference conditions.”

That last provision matters, and the honest reading is that it occupies the position this paper is about. A benchmarked profile is a result from an examination that a reader can set beside another result. It is scoped to one sector, it is a proposal rather than an operating program, and by the authors’ own account the value of its independent tier depends on an independence model, “assessor qualification, conflict

rules, sampling, oversight,” whose institutional design is “beyond the scope of this paper.” But it is not merely a vocabulary, and describing it as one would be wrong.

The January 2026 report goes further than conditions in two respects that matter here. It introduces four AI Assurance Levels, AAL-1 to AAL-4, “the higher levels of which provide greater confidence in audit findings,” and it treats recurrence directly: a mature auditing ecosystem, it argues, “will combine periodic deep assessments of slower-moving elements ... with event-triggered reviews of major changes ... and continuous automated monitoring,” a principle it summarizes as “Living assessments, not stale PDFs.”

Taken together, these two documents specify the conditions of an evaluation, propose the institutions that would accredit the evaluators, and propose a scale for the depth of assurance an audit provides. What the assurance levels calibrate is how much access and scrutiny an audit involves. That is a property of the examination, not a result of it.

3. The enforcement literature

A second body of work developed in parallel, and it is worth separating from the first because the two are often read as the same conversation.

The work described above concerns assessment. A different and larger effort concerns enforcement: what a system is permitted to do at the moment it acts. Through 2026 that effort converged on runtime control from several directions at once. One major vendor’s agent governance toolkit [5], released as open source, ships deny-by-default policy evaluation per tool call, per-agent identity, execution isolation, tamper-evident audit chaining, and conformance testing. The Coalition for Secure AI [3] and OWASP’s Agentic Security Initiative [10] published guidance along similar lines.

The convergence is itself informative. Several organizations, working separately and from different incentives, arrived at the same control point: the decision to act, evaluated before the action rather than after it. When independent parties converge on a location, the location is usually real.

4. Where the two literatures stop

What the enforcement literature produces is a control. What the assessment literature produces is a judgment about a control. These are different artifacts with different readers.

An enforcement framework specifies what a system does. An assessment framework specifies the conditions under which someone examines that system and reports on it. With one exception, noted above and scoped to a single sector, neither produces the third thing: a result from the examination that a reader can set beside another result and compare, across sectors, under conditions someone other than the participants defines.

That is not a criticism of either literature, and it would be a misreading to take it as one. The enforcement work is scoped to enforcement and says so. A runtime with a conformance suite does not claim to tell a procurement officer how one organization's governance compares to another's. The assessment work is scoped to conditions and says so. Each is doing the thing it set out to do.

The observation is about what sits between them. A conformance suite tells you whether a specific implementation passed a specific set of tests. A conditions checklist tells you whether a specific evaluation was conducted properly. A reader holding both still has two documents about one system, and no way to place that system relative to any other.

5. Three questions the published work leaves open

These are taken from the documents' own scope statements.

Recurrence. AEF-1 governs the conditions of a single evaluation exercise. It requires that the evaluator had adequate time, and it treats the evaluation as an event with a beginning and an end. For a static artifact that is the right scope. For a system whose behavior can change between the assessment and the deployment, a well-conducted one-time evaluation establishes a property of a system that no longer exists in that form. The January 2026 report identifies this directly and proposes periodic deep assessments, event-triggered reviews, and continuous automated monitoring in response. The proposal is specific about cadence and about what triggers a review. It is a proposal about how often to look, which is a different question from what the looking produces.

Enforcement work addresses this differently rather than solving it. Enforcement operates continuously by construction, which is a genuine difference from a point-in-time assessment. But continuous enforcement produces a continuous log, not a periodic measurement, and those are not the same artifact.

Comparability. The AEF-1 checklist records, for each condition, whether it was fulfilled and what the supporting evidence was. Two evaluators can each complete it rigorously, for two different systems, and produce two documents that cannot be set beside each other to answer which system was better governed. This is not an oversight. A conditions standard establishes that an evaluation was conducted properly; it does not produce a value on a common scale.

The enforcement side reaches the same boundary from the other direction. A conformance suite is pass or fail against its own tests. Two systems, each passing their respective suites, remain incomparable to a reader who has to choose between them.

Who verifies the attestation. The AEF-1 checklist is completed by the evaluator, about the evaluator, and published by the evaluator. Its conflict-of-interest conditions are substantive: the evaluator must publish a policy and apply it to the evaluation, and must separately disclose conflicts relevant to that evaluation. What the standard does not provide is an external party to confirm that a completed

checklist is accurate, and it does not accredit the evaluator who completed it. The January 2026 report proposes exactly that missing layer, an accreditation program and an auditor-of-auditors body with authority to revoke accreditation.

This should be stated carefully, because it is easy to misread as an accusation and it is not one. A self-completed attestation by a serious organization, published alongside its results, is a substantial improvement over the absence of one, and AEF-1 is explicit about its own scope: it establishes “an initial voluntary standard” for “a baseline set of operating conditions,” and it “does not cover all aspects of third-party AI evaluation.” The January 2026 report proposes exactly the accrediting institution AEF-1 does not specify. The question therefore shifts from whether an accreditation model has been proposed to whether the licensing regimes and standards bodies now forming will adopt comparable requirements.

6. One framing that should be retired

Binding at the moment of action rather than reviewing after it is no longer a contested position. Implementations and guidance converged on it through 2026. Any argument resting on enforcement-locus being novel should be retired, including arguments the author of this survey has previously made.

The convergence is evidence that the control point is correctly identified. It is not evidence that the comparability question has been answered.

Disclosure

The author is the founder of AAM Cyber, a cybersecurity consultancy, and has a commercial interest in the independent measurement of AI governance quality. The author also published *The Five Laws of AI Governance* [12], which addresses the same subject. Three concept crosswalks authored by the author are cataloged as Final Informative References in the NIST Online Informative References catalog [8], against the AI Risk Management Framework 1.0 [6], the Cybersecurity Framework 2.0 [7], and Special Publication 800-53 Revision 5, Release 5.2.0 [9]. Catalog inclusion is an informative reference, not an endorsement by NIST. The author has patent applications pending in this area.

References

[1] AI Evaluator Forum. *AEF-1: Minimum Operating Conditions for Independent Third Party AI Evaluations*. December 2025. https://aievaluatorforum.org/AEF_1_Minimum_Operating_Conditions_for_Independent_Third_Party_AI_Evaluations.pdf

- [2] Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., et al. *Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies*. January 2026. arXiv:2601.11699. <https://arxiv.org/abs/2601.11699>
- [3] Coalition for Secure AI (CoSAI), an OASIS Open project. *Secure Design Patterns for Agentic Systems* workstream. <https://www.coalitionforsecureai.org/> Accessed October 2026.
- [4] Hsu, M., Bankhwal, M., Moissinac, B., Werbach, K., Szpruch, L., and Hillenbrand, B. *Agent Reliability Profiles in Financial Services*. 2 October 2026. arXiv:2610.04123. <https://arxiv.org/abs/2610.04123>
- [5] Microsoft. *Agent Governance Toolkit*. MIT licence. <https://microsoft.github.io/agent-governance-toolkit/> Accessed October 2026.
- [6] National Institute of Standards and Technology. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. NIST AI 100-1. January 2023.
- [7] National Institute of Standards and Technology. *Cybersecurity Framework (CSF) 2.0*. NIST CSWP 29. February 2024.
- [8] National Institute of Standards and Technology. *Online Informative References (OLIR) Program: Informative Reference Catalog*. <https://csrc.nist.gov/projects/olir/informative-reference-catalog>
- [9] National Institute of Standards and Technology. *Security and Privacy Controls for Information Systems and Organizations*. NIST SP 800-53 Revision 5, Release 5.2.0. August 2025. <https://doi.org/10.6028/NIST.SP.800-53r5>
- [10] OWASP GenAI Security Project. *Agentic Security Initiative*. <https://genai.owasp.org/> Accessed October 2026.
- [11] Public Company Accounting Oversight Board. *Interim Ethics and Independence Standards, ET Section 101: Independence*. <https://pcaobus.org/oversight/standards/ethics-independence-rules/details/ET101>
- [12] Genece, P. *The Five Laws of AI Governance*. Zenodo, 2026. DOI 10.5281/zenodo.21271749. CC BY 4.0.