Network Working Group
Internet-Draft
Intended status: Informational
Expires: 14 November 2026

J. D. Hillier
Certisyn, Inc.

13 May 2026

AI Governance Verified — A Cryptographic Verification Standard for Agentic AI Governance in Regulated Industries

draft-hillier-certisyn-ai-governance-verified-00

Abstract

This document specifies a verification standard for the cryptographic attestation of agentic AI governance in regulated industries. It defines the Verification Reconciliation Object (VRO), the issuing-partner framework, the eight control areas through which AI governance posture is reconciled, three maturity-attestation levels (Documented, Operational, Adversarial-ready), and the cryptographic continuity requirements that together produce deterministic, independently reconstructable, auditor-grade attestations of agentic AI governance. The standard sits beneath ISO/IEC 42001:2023, the NIST AI Risk Management Framework, and other agentic AI governance frameworks, and produces the verifiable artefact those frameworks were designed to imply but do not deliver.

Status of This Memo

This Internet-Draft is submitted in full conformance with the provisions of BCP 78 and BCP 79.

Internet-Drafts are working documents of the Internet Engineering Task Force (IETF). Note that other groups may also distribute working documents as Internet-Drafts. The list of current Internet-Drafts is at https://datatracker.ietf.org/drafts/current/.

Internet-Drafts are draft documents valid for a maximum of six months and may be updated, replaced, or obsoleted by other documents at any time. It is inappropriate to use Internet-Drafts as reference material or to cite them other than as "work in progress."

This Internet-Draft will expire on 14 November 2026.

Copyright Notice

Table of Contents

1. Introduction

Agentic artificial intelligence is now operative across the workplace at a scale that exceeds the control envelope of every previously published governance framework. ISO/IEC 42001:2023 [ISO42001] specifies a management system for AI but does not produce a verifiable artefact. The NIST AI Risk Management Framework [NIST-AI-RMF] provides a functional taxonomy but issues no certification or attestation. The European Union AI Act [EU-AI-ACT] establishes obligations and prohibitions but leaves verification of compliance to national competent authorities and self-attestation.

No published standard issues a cryptographically anchored, deterministically reproducible monthly artefact of agentic AI governance.

This document closes that gap. It defines the verification artefact, the issuing-partner framework, the evidence requirements across eight control areas, the maturity-level attestation methodology, and the cryptographic continuity requirements that together produce a deterministic, auditor-grade AI governance attestation.

This document does not replace ISO/IEC 42001, the NIST AI Risk Management Framework, the EU AI Act, or any national framework. It sits beneath them and produces the artefact each was designed to imply but does not deliver. Where this document and any normative framework cited herein conflict on operational content, the cited framework prevails.

This document applies to organisations that operate, integrate, deploy, or expose AI systems — whether developed internally, procured from third-party model providers, or consumed via API. It does not specify AI model architecture, training procedure, or evaluation methodology. It specifies the verification of governance applied to AI use, not the AI itself.

2. Conventions and Definitions

The key words "MUST", "MUST NOT", "REQUIRED", "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "NOT RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC2119] [RFC8174] when, and only when, they appear in all capitals, as shown here.

For the purposes of this document, the following definitions apply.

Subject Entity: The organisation whose AI governance posture is the subject of verification.

AI System: Any deployed system, application, or service whose behaviour incorporates the use of one or more machine-learning models, irrespective of model architecture or training methodology.

Agentic System: An AI System that takes actions in the operating environment, generates outputs that influence downstream decisions, or operates with reduced or absent human-in-the-loop supervision.

Model Provider: An external party supplying a foundation model,

fine-tuned model, or model-as-a-service to the Subject Entity.

Deployment Context: The set of integrations, data sources, user populations, regulatory obligations, and risk attributes within which an AI System is operated.

Governance Surface: The composite of policies, controls, telemetry, evidence sources, and review cadences through which the Subject Entity governs the use of its AI Systems.

Verification Reconciliation Object (VRO): The deterministic, cryptographically anchored output of a conforming AI governance attestation under this standard.

Issuing Partner: A counterparty designated by the protocol operator to act as a co-issuer of VROs under this standard within a defined market or scope.

Attestation Period: The contiguous time interval over which a VRO asserts conformance.

Anchor Event: The cryptographic operation that binds a VRO to an immutable public settlement layer at issuance and at supersession.

Maturity Level: One of three levels (Documented, Operational, Adversarial-ready) defined in this document.

Supersession: The lifecycle event by which a new VRO replaces a prior VRO.

3. Architectural Overview

Conforming attestations under this standard are produced by a verification infrastructure organised as five architectural components. Internal design, scoring methodology, and calibration logic are not in scope for this document.

The Evidence Ingestion and Normalization Layer accepts AI governance Evidence Artefacts from Subject Entity systems — model registries, identity-provider telemetry, prompt and output logs, sanctioned-application controls, incident records — and normalises representation across heterogeneous source formats.

The Reconciliation Confidence Engine reconciles Conformance Claims against normalised evidence and produces a deterministic reconciliation output.

The Verification State Machine maintains the lifecycle state of each VRO through intake, evidence ingestion, reconciliation, anchoring, issuance, supersession, and revocation.

Entity Graph Propagation propagates verification state across related entities where continuity is in scope.

The Attestation Protocol produces the final VRO, performs the Anchor Event, registers the artefact in the public attestation registry, and binds the issuing-partner identity.

Conforming attestations are deterministic. Given the same Conformance Claims and the same Evidence Artefacts processed through the same Attestation Protocol version, the same VRO SHALL be produced. Determinism applies to the verification operation, not to the AI systems being verified.

4. AI Governance Verification Requirements

The following subsections specify the eight control areas through which agentic AI governance is reconciled under this standard.

4.1. Area 1: AI inventory and shadow-AI discovery

Subject Claim: The Subject Entity maintains a current inventory of AI Systems in operation across its workforce, integrations, and infrastructure, and detects use of unsanctioned or undisclosed AI Systems within its operating environment.

Evidence Categories: AI System register; identity-provider telemetry of AI service authentications; endpoint or network telemetry of model API egress; sanctioned-application register; shadow-AI detection output; periodic reconciliation reports.

Verification Expectation: Reconciliation of declared inventory against detected use across the Attestation Period; exception cases reconciled against the disclosure or remediation register.

Anchor Requirement: Anchored at Attestation Period start and end. Material expansions of inventory require a supersession anchor.

4.2. Area 2: Use-case classification and risk assessment

Subject Claim: The Subject Entity classifies each AI System by use-case category and risk tier, and records the classification together with the rationale and the residual-risk position.

Evidence Categories: Use-case classification register; risk

assessment artefact for each AI System; risk-tier policy artefact; review records evidencing periodic reassessment; exception register for ungoverned use cases.

Verification Expectation: Reconciliation of declared classifications against the policy taxonomy; reconciliation of risk tier against deployment context evidence; identification of classification drift over the Attestation Period.

Anchor Requirement: Anchored at issuance; supersession on material change in deployment context, risk tier, or use-case scope.

4.3. Area 3: Model and data provenance

Subject Claim: The Subject Entity records and maintains provenance evidence for each AI System in use, including model identity, model version, Model Provider identity, training data disclosures (where available), and update or fine-tuning lineage.

Evidence Categories: Model registry entries with provider, version, and lineage data; Model Provider transparency reports or model cards where supplied; training data attestations where available; fine-tuning records; vendor change log.

Verification Expectation: Reconciliation of recorded provenance against AI System operational state; reconciliation of fine-tuning lineage against change-management evidence; identification of model-version drift across the Attestation Period.

Anchor Requirement: Anchored at issuance and at each material model-version change or Model Provider change.

4.4. Area 4: Sanctioned-application control

Subject Claim: The Subject Entity restricts use of AI Systems to those that have been explicitly sanctioned for the applicable user population and use-case context, consistent with the asserted Maturity Level.

Evidence Categories: Sanctioned-application allow-list policy; endpoint or network enforcement evidence; exception register; user-population scope evidence; periodic review records.

Verification Expectation: Reconciliation of declared allow-list against enforcement telemetry; reconciliation of exception cases against the exception register; identification of unsanctioned use over the Attestation Period.

Anchor Requirement: Anchored at issuance and on material allow-list changes.

4.5. Area 5: Prompt and output governance

Subject Claim: The Subject Entity governs the content of prompts submitted to AI Systems and outputs produced by AI Systems, including controls preventing disclosure of sensitive data to external AI Systems and controls preventing high-risk output content from entering downstream processes.

Evidence Categories: Prompt-content policy artefact; output-content policy artefact; prompt-monitoring telemetry; output-review telemetry; data-loss-prevention rules applied to AI traffic; high-risk content exception register.

Verification Expectation: Reconciliation of declared content controls against monitoring telemetry; reconciliation of exception handling against review evidence; identification of control bypass or drift over the Attestation Period.

Anchor Requirement: Anchored at issuance and on each material change in control scope, sensitive-content taxonomy, or enforcement state.

4.6. Area 6: Identity and access control for AI

Subject Claim: The Subject Entity enforces identity and access controls on the use of AI Systems consistent with the asserted Maturity Level, including authentication, authorisation, multi-factor enforcement, and segregation between human and machine principals.

Evidence Categories: Identity-provider telemetry for AI service authentications; access assignment register for AI Systems; multi-factor enrolment coverage report; service-account inventory; access review records.

Verification Expectation: Reconciliation of declared access model against assignment register; reconciliation of multi-factor enforcement against identity-provider evidence; reconciliation of service-account use against the inventory and policy.

Anchor Requirement: Anchored at issuance and at the conclusion of each scheduled access review cycle.

4.7. Area 7: Logging, telemetry, and auditability

Subject Claim: The Subject Entity captures, retains, and protects logs of AI System use sufficient to permit retrospective reconstruction of governance-relevant events, with retention and integrity properties consistent with the asserted Maturity Level.

Evidence Categories: Logging policy artefact; log content schema; log retention configuration; log-integrity attestation; access-control evidence for log stores; sampling or audit records.

Verification Expectation: Reconciliation of declared logging scope against captured content; reconciliation of asserted retention against log-store configuration; reconciliation of asserted integrity properties against evidence of log-store immutability or chain protection.

Anchor Requirement: Anchored at issuance and on material changes to logging scope, retention, or integrity configuration.

4.8. Area 8: Incident, drift, and escalation response

Subject Claim: The Subject Entity operates an incident-response capability for AI-related events, including hallucination, output failure, prompt-injection, data exfiltration, model drift, and high-impact misuse, with defined escalation paths and post-event review consistent with the asserted Maturity Level.

Evidence Categories: AI incident-response policy; incident register; incident classification taxonomy; escalation records; post-event review reports; remediation evidence; drift-monitoring telemetry.

Verification Expectation: Reconciliation of declared response capability against the incident register over the Attestation Period; reconciliation of escalation evidence against the declared escalation paths; reconciliation of remediation evidence against committed actions.

Anchor Requirement: Anchored at issuance and at the conclusion of each material incident response or post-event review cycle.

5. Maturity Level Attestation

A conforming VRO under this standard SHALL attest a Maturity Level for each of the eight control areas. Different control areas MAY attest at different Maturity Levels within a single VRO; the overall VRO attestation is the minimum Maturity Level attested across the eight control areas unless otherwise asserted.

Maturity Level Documented reflects the baseline expectation that governance content exists, that policy artefacts are written, that an inventory is maintained, and that a designated owner is accountable. Evidence requirements at this level emphasise the existence of declared content and basic operational artefacts over an Attestation Period of at least three (3) consecutive months.

Maturity Level Operational reflects the expectation that controls are not only documented but exercised: telemetry collected, periodic reviews occurring, exceptions recorded and handled, and the governance surface responding to material change. Evidence requirements add depth of telemetry, review-cadence evidence, exception handling, and continuity over an Attestation Period of at least six (6) consecutive months.

Maturity Level Adversarial-ready reflects the expectation that governance withstands adversarial conditions: prompt-injection attempts, model-drift events, data-exfiltration attempts via AI channels, sophisticated misuse, and dependency failures at the Model Provider. Evidence requirements add continuity, defence-in-depth evidence, red-team or adversarial evaluation attestation where in scope, and continuous reconciliation over an Attestation Period of at least twelve (12) consecutive months.

A Subject Entity that progresses to a higher Maturity Level for any control area SHALL be issued a superseding VRO recording the progression. The prior VRO is preserved and marked as superseded.

6. Verification Reconciliation Object (VRO)

A conforming VRO under this standard SHALL contain, at minimum:

* Subject Entity identifier.
* Attestation Period start and end timestamps.
* Maturity Level attested for each of the eight control areas.
* Conformance Claims as asserted by the Subject Entity.
* Evidence categories ingested and reconciliation outcome for each.
* AI System inventory snapshot at Attestation Period end.
* Issuing Partner identity and seat designation.
* Anchor Event identifiers binding the VRO to the public settlement layer.

* Verification State Machine state at issuance.

* Supersession chain reference, where applicable.

* Conformance statement of this standard, version 1.0.

A VRO MAY be revoked by the Issuing Partner upon determination of material non-conformance, evidence falsification, undisclosed incidents, or other circumstances rendering the original attestation unreliable. Revocation does not delete the VRO; it records a revocation state, the revocation reason class, and the Anchor Event binding the revocation to the public settlement layer.

Each issued VRO SHALL be registered in the public attestation registry.

7. Issuing Partner Requirements

An organisation seeking designation as an Issuing Partner under this standard SHALL demonstrate, at minimum:

* Operational capacity to assess AI governance posture across the eight control areas at the Maturity Level for which issuance is sought.

* Demonstrable competence in AI deployment models, identity and access controls, prompt and output monitoring, and incident response.

* Independence from the Subject Entity at the engagement level, with declared conflicts of interest disclosed and managed.

* Independence from any Model Provider whose models are within the scope of attestation, or, where dependency exists, declared and managed under a stated independence protocol.

* Adherence to the protocol operator’s Partner Code of Conduct.

* Acceptance of the Designation Schedule terms applicable to the relevant market and seat.

An Issuing Partner SHALL NOT, for a given Subject Entity engagement, simultaneously act as the implementing vendor, deployment integrator, or operator of the AI Systems being verified.

8. Cryptographic Continuity Requirements

Each VRO SHALL be cryptographically anchored to an immutable public settlement layer at the Anchor Event. The hash committed at the Anchor Event SHALL be a one-way function of the VRO content, Issuing Partner identity, and timestamp, computed under a digest algorithm of at least 256-bit strength.

A VRO issued under this standard SHALL remain a conforming artefact across regulatory regime changes occurring within or after the Attestation Period.

The Anchor Event binding SHALL remain independently verifiable in the event of a Model Provider ceasing to operate, withdrawing a model, or being acquired or restructured. VROs issued during the operating life of a withdrawn model are not retroactively invalidated.

9. Standards Alignment

This standard is interoperable with adjacent frameworks. Conforming VROs MAY be referenced within audit, certification, and regulatory artefacts produced under:

ISO/IEC 42001:2023 [ISO42001]: AI management-system controls map to the eight control areas in this document.

NIST AI Risk Management Framework [NIST-AI-RMF]: Functions (Govern, Map, Measure, Manage) map to evidence categories within the eight control areas.

EU AI Act [EU-AI-ACT]: Provider and deployer obligations MAY be evidenced through conforming VROs where the obligation is verifiable through reconcilable evidence. The EU AI Act remains authoritative for legal compliance determinations.

ISO/IEC 27001:2022 [ISO27001]: AI control areas intersecting information security are interoperable with ISO 27001 Annex A controls.

Essential Eight Verified [I-D.hillier-certisyn-essential-eight-verified] cross-references application control, privilege restriction, and authentication evidence categories.

10. Conformance

An attestation artefact MAY claim conformance to this standard if and only if it satisfies every requirement specified in this document. Partial conformance is not recognised. Variant conformance to a subset of control areas without the full eight-area scope is not recognised.

The public attestation registry constitutes the authoritative record of issued VROs.

11. IANA Considerations

This document has no IANA actions.

12. Security Considerations

Agentic AI governance operates under adversarial conditions distinct from traditional cybersecurity. Implementations of this standard SHOULD pay particular attention to prompt-injection resistance, model-drift detection, and exfiltration paths through AI channels that may bypass traditional data-loss-prevention controls.

Issuing Partners are required to be independent from the Subject Entity and from Model Providers whose models are within attestation scope.

The Anchor Event binding SHOULD use a digest algorithm of at least 256-bit strength and a public settlement layer with no single private operator capable of extinguishing the binding.

This standard does not address the correctness or safety of the AI Systems being governed. It addresses the verifiability of governance applied to those systems.

13. References

13.1. Normative References

[RFC2119] Bradner, S., "Key words for use in RFCs to Indicate Requirement Levels", BCP 14, RFC 2119, DOI 10.17487/RFC2119, March 1997, <https://www.rfc-editor.org/rfc/rfc2119>.

[RFC8174] Leiba, B., "Ambiguity of Uppercase vs Lowercase in RFC 2119 Key Words", BCP 14, RFC 8174, DOI 10.17487/RFC8174, May 2017, <https://www.rfc-editor.org/rfc/rfc8174>.

13.2. Informative References

[EU-AI-ACT]
Union, E., "Regulation (EU) 2024/1689 — Artificial Intelligence Act", 2024.

[I-D.hillier-certisyn-essential-eight-verified]
Hillier, J. D., "Essential Eight Verified — A Cryptographic Verification Standard for the ACSC Essential Eight Maturity Model", May 2026.

[ISO27001] Standardization, I. O. for., "Information security, cybersecurity and privacy protection — Information security management systems — Requirements", ISO/IEC 27001:2022, 2022.

[ISO42001] Standardization, I. O. for., "Information technology — Artificial intelligence — Management system", ISO/IEC 42001:2023, 2023.

[NIST-AI-RMF]
Technology, N. I. of S. and., "AI Risk Management Framework", 2023.

Acknowledgments

The author thanks the ANZ Founding Partner cohort for early review of this draft.

Author’s Address

Joel David Hillier
Certisyn, Inc.
Email: jhillier@certisyn.com
URI: https://certisyn.com/