OVERT

VERSION 1.0

OBSERVABLE VERIFICATION EVIDENCE FOR RUNTIME TRUST

An Open Standard for Runtime Trust in AI Systems

Initial Release — Version 1.0

DATE	March 2026
EDITORS	GLACIS Technologies, Inc.
CONTACT	overt-review@glacis.io
STANDARD URL	overt.is
IPR POLICY	overt.is/ipr-policy

OVERT defines how runtime controls, boundary-enforcement decisions, measurement outputs, and response actions are bound to observable verification evidence that independent parties can validate without requiring protected-content egress. It serves both as infrastructure for verifiable AI governance and as a runtime trust layer for prevention, detection, containment, and post-incident reconstruction.

Informative Foreword: The Verification and Security Gap

Existing AI governance frameworks specify what controls should exist. They generally do not specify how to produce independent, tamper-evident proof that those controls executed on a given interaction, under a given configuration, at a given time. That gap is not only a governance verification gap. It is also an AI security visibility and containment gap.

As AI systems move into consequential and adversarial settings, operators and relying parties increasingly need more than policy documents, self-generated logs, and periodic audit narratives. They need trustworthy evidence of what executed at the runtime boundary; what traffic was within mediation scope; what was allowed, denied, sampled, escalated, or overridden; whether the enforcing component was the expected one; and whether those records can be verified independently after an incident without creating a new protected-content disclosure channel.

Current practice often lacks five properties that mature security operations require:

- **Trusted execution evidence** showing which enforcing component and configuration were active when a governed action occurred.
- **Reliable runtime coverage accounting** showing what traffic and action classes were in scope, what was excluded, and how denominators were derived.
- **Tamper-evident telemetry** that is not reducible to operator-controlled logs.
- **Independent verification of enforcement events** including permit, deny, override, escalation, and response actions.
- **Post-incident reconstruction without routine content disclosure** so relying parties can verify event history without turning attestation into a new protected-data egress channel.

OVERT addresses that narrower problem. It does not adjudicate the merits of any particular governance, security, or legal dispute. It specifies how to produce independently assessable records of control execution, measurement, and response without requiring protected-content egress.

Keywords

The key words "SHALL", "SHALL NOT", "SHOULD", "SHOULD NOT", "RECOMMENDED", "MAY", and "OPTIONAL" in this document are to be interpreted as described in BCP 14 [RFC 2119] [RFC 8174] when, and only when, they appear in all capitals, as shown here.

Requirements marked SHALL are normative and required for conformance. Requirements marked SHOULD or RECOMMENDED are normative recommendations. Requirements marked MAY indicate

permitted behavior that is truly optional for conformance. All annexes (A through F) are informative. All notes, examples, architectural references, and case studies are informative.

Table of Contents

Intellectual Property Rights Notice

Patent Covenant. GLACIS Technologies irrevocably commits not to assert patent claims against any implementation that conforms to this standard, regardless of which registered Protocol Profile the implementation uses, whether the implementation is commercial or non-commercial, and whether the implementation combines the attestation capabilities defined here into a unified pipeline or implements them separately. This covenant is irrevocable, runs with the patent, and is binding on GLACIS Technologies and its successors and assigns.

Patent Disclosures. GLACIS Technologies holds patent filings related to certain methods described in this standard. Full patent disclosures, claim-scope details, and guidance for alternative implementations are published at overt.is/ipr-policy.

Standard Normative Requirements. The normative requirements of this standard — including the attestation envelope structure, conformance levels, auditor verification procedures, and agentic governance controls — define functional properties and interoperability requirements. They do not mandate the use of any specific patented method. Conformance with this standard can be achieved

through multiple architectural approaches, and implementers are free to select cryptographic constructions and architectural patterns that satisfy the normative requirements.

Contributor Disclosure. Contributors to this standard who are aware of potentially essential patent claims are expected to disclose them. No single commercial entity has unilateral control over protocol profile registration, conformance criteria, or certification governance. The Protocol Profile Registration Process is governed by the criteria in Section 22.6. Multiple profiles are permitted. Conformance requires exactly one registered profile per deployment.

PART 1: FOUNDATIONS

1. Purpose and Scope

1.1 Purpose

OVERT defines an open standard and certification framework for attested AI runtime control systems. It specifies requirements for generating, storing, preserving, and verifying cryptographic proof that declared governance and runtime control decisions executed under a defined configuration, within a bounded time interval, without requiring protected-content egress.

In this role, OVERT serves three related purposes. First, it supports verifiable AI governance by making policy execution, oversight actions, measurement outputs, and response activities independently assessable. Second, it provides a low-level control-and-evidence substrate for AI runtime security by enabling attested runtime identity, policy-mediated execution decisions, evidence that declared tool and boundary controls executed within the attested scope, tamper-evident telemetry, attested response actions, and post-incident reconstruction of control execution history. Third, it provides the conformance, independence, and assessment model by which relying parties can distinguish self-asserted deployment claims from independently assessed, evidence-grade runtime mediation.

OVERT is not itself a universal runtime-control product. Enforcement is performed by conformant arbiters, sidecars, gateways, proxies, or equivalent runtime-control implementations operating under a registered Protocol Profile. OVERT defines what those implementations SHALL prove, what evidence an independent attestation provider SHALL verify, and what a qualified assessor SHALL examine when a conformance claim is made.

OVERT does not replace governance frameworks, security engineering disciplines, runtime-control implementations, or legal analysis. Organizations remain responsible for defining policies, selecting controls, securing infrastructure, evaluating models, and satisfying applicable law. OVERT specifies how to produce temporally bound, tamper-evident, independently verifiable artifacts demonstrating that declared controls executed and that attested measurements and response actions can be reconstructed and checked by relying parties.

OVERT attests control execution and associated evidence quality. It does not attest the truthfulness of model outputs, the absence of hallucination, the absence of compromise, or the adequacy of the operator's policies. Attestation artifacts are designed to support authenticity, integrity, timing, auditability, and chain of custody. Their legal relevance, admissibility, and sufficiency remain questions of applicable law and context.

1.2 Limitations of Attestation

OVERT does not:

- Replace endpoint, cloud, network, application, platform, model, or software supply-chain security controls.
- Detect every attack, abuse path, or failure mode by itself.
- Guarantee that declared policies are adequate, lawful, or well configured.
- Make an unsafe, insecure, or poorly governed AI system safe merely because attestations are produced.
- Attest the quality, accuracy, truthfulness, fairness, robustness, or cybersecurity of model outputs as substantive properties.
- Eliminate the need for human incident response, forensic investigation, sector-specific controls, or domain-specific validation.
- Guarantee legal compliance, regulatory approval, evidentiary admissibility, or insurance coverage.
- Prove the absence of compromise, data poisoning, prompt-injection success, or unauthorized access outside the attested scope.
- Substitute attestation artifacts, managed deployment claims, or certification language for actual in-path runtime mediation.
- Treat operator or vendor assertions about deployment completeness as sufficient for AAL-4 or Level 4 conformance absent the independence requirements of this standard.

OVERT proves, within the claimed scope and assurance level, that certain controls, measurements, and response actions were executed or recorded in the manner specified by this standard. Whether those controls were sufficient for a given use case remains a separate question.

Training-time operations (data preparation, model training, experiment tracking, fine-tuning), data lifecycle management (versioning, freshness, deletion), and platform infrastructure security (vulnerability management, SDLC, patching, secrets management) are outside the OVERT attestation scope. These are important controls addressed by frameworks including DASF, NIST SP 800-53, and ISO 27001. OVERT complements but does not replace them.

A future OVERT Build Assurance Profile may define attestation requirements for training, data, and platform lifecycle controls. Until such a profile is published, implementers SHOULD NOT represent OVERT conformance as covering training, data lifecycle, or platform infrastructure security. Conformance statements that could be misread as covering these surfaces are non-conformant with the spirit of this standard.

1.3 Scope

This standard applies to:

- **AI system operators** deploying AI in regulated industries (healthcare, financial services, insurance, employment, education, housing)
- **AI system developers** building products subject to governance obligations
- **Security teams, incident responders, auditors, procurement reviewers, and regulators** who need to verify control execution without routine access to protected content
- **Insurers** who need quantitative, cryptographically verifiable data to price AI risk
- **Agentic AI systems** where autonomous agents execute tool calls, access external resources, and make decisions without step-by-step human oversight

1.4 Relationship to Existing Standards

OVERT operates beneath and alongside existing AI governance, security, attestation, and certification frameworks. Its role is to provide a trust, execution-control, telemetry, evidence, and conformity-assessment substrate for AI systems: a mechanism by which governance and security-relevant events can be bound to runtime state, recorded without protected-content egress, and independently verified.

OVERT therefore complements, but does not replace, governance frameworks such as NIST AI RMF and ISO/IEC 42001; security frameworks such as NIST SP 800-53, FedRAMP, and zero-trust architectures; attestation architectures such as IETF RATS; and implementation products that actually mediate runtime actions. Those frameworks and products specify objectives, controls, management processes, trust relationships, or execution mechanisms at broader organizational and system levels. OVERT specifies how to generate and verify cryptographic records of declared control execution for AI systems and agentic workflows, and how those claims are assessed for conformance.

Conformance with OVERT is not a determination of compliance with any other standard, law, or regulatory regime. Rather, OVERT artifacts may support evidence for requirements defined elsewhere, subject to the scope, assurance level, and limitations of this standard.

OVERT operates beneath and is complementary to:

Standard	Role	OVERT Relationship
NIST AI 100-1 (AI RMF 1.0)	Risk management functions	OVERT provides attestation artifacts supporting evidence that GOVERN/MAP/MEASURE/MANAGE activities were executed
ISO/IEC 42001:2023	AI management system	OVERT supports evidence for A.6.2.8 (event logging) and extends event records to tamper-evident, third-party-verifiable attestation
EU AI Act (Regulation 2024/1689)	Regulatory requirements	OVERT supports evidence for Article 12 (automatic logging) and aspects of Article 9 (risk management) documentation requirements. Regulation (EU) 2024/1689 generally applies from 2 August 2026. Article 6(1) and corresponding obligations apply from 2 August 2027. Annex III systems (Article 6(2)) follow the general application date
IETF RATS (RFC 9334)	Remote attestation architecture	OVERT instantiates the Attester/Verifier/Relying Party model for AI attestation
NIST OSCAL	Machine-readable compliance	OVERT attestation packs are expressible as OSCAL assessment results
Registered OVERT Protocol Profile	Implementation specification	Specifies cryptographic constructions, envelope schemas, key derivation, and signal formats implementing this standard. Protocol Profile 1.0 is the initial registered profile; see Annex B
NIST SP 800-53 Rev 5	Security and privacy controls	OVERT maps to AU (Audit), SI (System Integrity), IA (Identification/Authentication) families
FedRAMP Moderate Baseline	Federal cloud authorization	OVERT attestation architecture supports evidence for FedRAMP AU and SI control families
NIST AI RMF GenAI Profile	GenAI-specific guidance	OVERT provides attestation artifacts for GenAI-specific GOVERN/MEASURE/MANAGE recommendations
NIST SP 800-207	Zero Trust Architecture	OVERT trust architecture is complementary; "untrusted SUT" model is distinct from ZTA network assumptions

Standard	Role	OVERT Relationship
OMB M-25-21 / M-25-22	Federal AI procurement	OVERT attestation packs support AI use case inventory and risk management documentation requirements. M-25-22 applies to solicitations issued on or after October 1, 2025 (180 days after issuance). Agency AI inventory/reporting obligations are in M-25-21 (initial reporting targeting April 2026 per OMB instructions). M-25-22 excludes National Security Systems

1.5 Design Principles

1. **Attestation by construction, not assertion.** Controls produce cryptographic proof as a byproduct of execution, not as a separate documentation exercise.
2. **Privacy by architecture, not policy.** Protected content never leaves the operator's environment. Only cryptographic commitments cross trust boundaries.
3. **Independence by structure.** The entity attesting to governance is structurally independent of the entity being governed. Self-attestation is not compliant.
4. **Statistical rigor by default.** Safety claims carry confidence intervals, sample sizes, and auditor-reproducible methodologies. Unquantified assertions are not attestation artifacts.
5. **Open by design.** This standard is open for implementation by any party. Reference implementations are open-source.
6. **Security-supporting evidence by observation.** The architecture that produces governance evidence occupies the same inline position, binary identity measurement, behavioral monitoring, and tamper-evident recording paths that security detection requires. Within the attested scope, OVERT produces security-supporting evidence — not a complete security architecture. Whether that evidence is sufficient for a given security objective depends on mediation scope, denominator independence, arbiter isolation, IAP topology, and the operator's broader security posture.

2. Normative References

The following documents are referenced normatively within this standard:

2.1 Normative References

- RFC 2119 / RFC 8174: Key words for use in RFCs to Indicate Requirement Levels (BCP 14)
- NIST AI 100-1: AI Risk Management Framework 1.0 (January 2023)
- ISO/IEC 42001:2023: Information Technology — Artificial Intelligence — Management System
- ISO/IEC 22989:2022: Artificial Intelligence — Concepts and Terminology
- RFC 9334: Remote ATtestation procedureS (RATS) Architecture
- RFC 6962: Certificate Transparency
- EU Regulation 2024/1689: AI Act
- NIST SP 800-207: Zero Trust Architecture
- NIST SP 800-53 Rev 5: Security and Privacy Controls for Information Systems and Organizations
- A registered OVERT Protocol Profile (see Annex B for Protocol Profile 1.0, the initial registered profile)

2.2 Informative References

- RFC 8949: Concise Binary Object Representation (CBOR) — Section 4.2, Deterministic Encoding (used by Protocol Profile 1.0)
- RFC 5869: HMAC-based Extract-and-Expand Key Derivation Function (HKDF) (used by Protocol Profile 1.0)
- RFC 8785: JSON Canonicalization Scheme (JCS) (used by Protocol Profile 1.0)
- NIST SP 800-208: Recommendation for Stateful Hash-Based Signature Schemes
- FIPS 204: Module-Lattice-Based Digital Signature Standard (ML-DSA)
- FIPS 205: Stateless Hash-Based Digital Signature Standard (SLH-DSA)
- OWASP Top 10 for Agentic Applications (December 2025)
- NIST AI RMF Generative AI Profile (July 2024)
- Colorado SB 24-205: Consumer Protections for Artificial Intelligence (effective June 30, 2026, as amended by SB 25B-004)
- OMB Memorandum M-25-21: Accelerating Federal Use of AI through Innovation, Governance, and Public Trust
- OMB Memorandum M-25-22: Driving Efficient Acquisition of Artificial Intelligence in Government
- RFC 9711: Entity Attestation Token (EAT)
- RFC 9162: Certificate Transparency Version 2.0
- AIUC-1: AI Use Case Inventory and Classification (January 2026)

> ***Note:*** *Protocol Profiles SHOULD include a documented post-quantum cryptographic transition plan referencing NIST FIPS 204 (ML-DSA) or FIPS 205 (SLH-DSA). The informative references to FIPS 204 and FIPS 205 above are included to facilitate such planning.*

3. Terms and Definitions

For the purposes of this standard, the terms in ISO/IEC 22989:2022 and the following apply:

3.1 attestation: A cryptographically signed statement by an independent notary that a specific governance action occurred, at a specific time, under a specific system configuration.

3.2 arbiter: An enforcement component deployed at the boundary between an AI system and external resources that intercepts, evaluates, and gates actions against defined policy.

3.3 co-epoch binding: The cryptographic linkage of an attestation to the exact binary identity and network isolation state of the system during a bounded time interval (epoch).

3.4 digest publication ledger (DPL): A per-epoch publication of request commitments enabling third-party verification of sampling completeness.

3.5 epoch: A bounded time interval during which system configuration is attested as stable by the notary network. Duration is configurable; recommended values are specified in the registered Protocol Profile.

3.6 attestation assurance level (AAL): One of four tiers (AAL-1 through AAL-4) describing the cryptographic verifiability and independence of governance attestation artifacts. See Section 4.1.

3.7 non-egress attestation: An attestation generation architecture in which protected content never leaves the operator's environment; only cryptographic commitments cross trust boundaries.

3.8 provisional receipt: A locally-signed attestation generated synchronously during enforcement, pending asynchronous counter-signature by the notary network.

3.9 receipt: A cryptographic artifact proving that a specific enforcement decision was made, at a specific time, under a specific configuration, and attested by an independent party.

3.10 statistical safety signal: A quantified statement of the form "with [confidence]% confidence, the violation rate for [policy] did not exceed [bound]% during [epoch]," derived from cryptographically verifiable random sampling.

3.11 tool call: An action by an AI agent that invokes an external capability — API call, database query, file operation, code execution, communication, or any interaction with systems outside the model's internal computation.

3.12 human-in-the-loop (HITL) interaction: Any event where a human provides consent, approval, review, correction, override, or other governance-relevant input to an AI system workflow. HITL interactions are attestable events subject to the same attestation requirements as automated enforcement decisions.

3.13 notary network: A set of geographically distributed, structurally independent nodes that collectively validate attestations, requiring agreement of t-of-n nodes before a valid receipt can be issued. No single node can forge or suppress attestation artifacts. The signature construction achieving the t-of-n property (threshold signature, multi-signature, or other scheme) is specified in the registered Protocol Profile.

3.14 independent attestation provider (IAP): An entity structurally independent of the AI system operator that operates notary infrastructure, validates attestations, and publishes transparency log entries.

3.15 protocol profile: A registered implementation specification defining cryptographic constructions, envelope schemas, key derivation methods, and receipt formats that implement this standard. Multiple profiles may coexist. Conformance requires exactly one registered profile per deployment.

3.16 mediation scope statement: A signed declaration identifying the action types, components, tenants, and traffic paths covered by the attestation system. Published in machine-readable form and referenced in risk signal computation. The mediation scope statement defines what is "in scope" for coverage ratio, exposure window, and other signal denominators.

3.17 qualified risk officer: An individual with documented authority and competence to make risk classification and severity determination decisions under GOV-3. Competence criteria are defined by the operator's risk management policy and SHALL include documented training in AI risk management. The qualified risk officer for an AI system SHALL NOT be the system's sole developer. Referenced in GOV-3.5 as the required policy artifact signer.

3.18 baseline intent declaration: A machine-readable, versioned, hash-chained governance artifact specifying the permitted agent topology, behavioral bounds per agent class, permitted spawn relationships, model bindings, and human oversight checkpoints for a deployment. Published to the transparency log. The baseline intent declaration is the reference artifact against which behavioral drift (3.21) is measured.

3.19 graph complexity metric: A quantitative measure of agentic execution topology — including edge count, branching factor, and depth utilization — computed per execution and evaluated relative to thresholds declared in the baseline intent declaration (3.18).

3.20 causal drift attribution: The process of tracing a detected behavioral drift signal in one agent to a correlated change in an upstream agent via parent-child attestation linkages in the transparency log.

3.21 behavioral drift: A statistically significant change in an agent's output distribution, tool selection distribution, or interaction patterns that occurs within authorized behavioral bounds — distinct from a policy violation. Behavioral drift is detected by sequential statistical methods operating on measurement features produced by the evaluation instrument specified in the registered Protocol Profile.

4. Architecture Overview

OVERT architecture defines the trust model by which AI governance claims and AI runtime security claims can be made independently assessable. The architecture is designed to answer a bounded set of questions that existing governance documentation and operator-controlled logs answer poorly: what component enforced the decision, what policy state and network state were in effect, what event occurred at the boundary, what was measured or escalated, and whether those records can be verified without trusting the system under test.

The OVERT architecture intentionally separates four roles that are often collapsed in market messaging: the standard defines the normative requirements, runtime-control implementations mediate execution, independent attestation providers verify attestations and operate notary infrastructure, and qualified assessors certify conformance claims at the levels this standard requires. A single commercial offering MAY package more than one operational role, but packaging does not relax the independence requirements stated in this standard.

The architectural relationship to security is positional, not comprehensive. NGAV and EDR shifted endpoint security toward runtime behavior, policy-mediated execution, tamper-evident telemetry, containment, and post-incident reconstruction. OVERT occupies an analogous inline position for AI systems and produces security-supporting evidence within the attested scope: attested runtime identity, policy-mediated tool and boundary-control decisions, tamper-evident telemetry, inter-agent trust controls, capability mediation records, evidence-preserving response, and verification without routine protected-content egress (Design Principle 6). OVERT is not a complete security product. It does not by itself prove that every declared boundary was complete or uncompromised, does not establish comprehensive defense, and does not guarantee that mediation scope covers all security-relevant traffic. The attestation infrastructure it specifies produces security-supporting evidence within the declared scope; whether that evidence is sufficient for a given security objective

depends on scope completeness, denominator independence, arbiter isolation, IAP resilience, and the operator's broader security controls.

4.1 Attestation Assurance Levels (AAL)

OVERT defines four attestation assurance levels. Each level subsumes the requirements of all lower levels. The levels represent increasing degrees of verifiability: AAL-1 and AAL-2 provide documentation and process records suitable for policy declaration and organizational governance; AAL-3 adds machine-generated telemetry and measurement outputs that can be operationally useful for monitoring, but that remain operator-controlled; and AAL-4 adds independently verifiable, cryptographically bound runtime evidence of enforcement, measurement, and response events. Higher AAL tiers produce stronger evidence within the attested scope; they do not by themselves establish comprehensive security.

Level	Name	Description	Verification Model
AAL-1	Policy Documentation	Written governance policies exist	Self-asserted; manual review
AAL-2	Process Records	Operational records of governance activities exist	Self-attested; auditor must trust operator
AAL-3	Automated Monitoring	System generates continuous governance telemetry	Machine-generated but operator-controlled
AAL-4	Cryptographic Attestation	Independent third party produces tamper-evident proof of control execution	Third-party verifiable; zero content access required

OVERT conformance requires AAL-4 attestation for all controls designated as AAL-4 in this standard. Controls designated AAL-1, AAL-2, or AAL-3 require the specified level. Conformance is assessed per-control, not globally.

AAL-1 through AAL-3 remain valid for organizational governance activities (policy drafting, training, culture) where cryptographic attestation is not architecturally applicable. But for any control that involves runtime AI system behavior — enforcement, monitoring, logging, incident detection — AAL-4 is mandatory.

4.1.1 Deployment Architecture and AAL Mapping

The following table maps deployment architectures to the maximum attestation assurance level achievable under each architecture. The mapping is normative.

Deployment Architecture	Maximum AAL	Rationale
No attestation infrastructure	AAL-2	Operator-generated records only; no independent verification

Deployment Architecture	Maximum AAL	Rationale
Single notary, operator-controlled	AAL-3	Independent attestation present but operator controls the notary
Single notary with hardware-rooted measurement (TEE), operator-controlled	AAL-3	Hardware root of trust strengthens measurement; operator still controls the notary
Multiple notaries (t-of-n), single operating entity	AAL-3	Multi-notary verification present but organizational independence not met
Single notary, independent third party (IAP)	AAL-4	Independent attestation with third-party trust root
Multiple notaries (t-of-n), independent operating entities	AAL-4	Highest assurance; multi-entity independence with third-party verifiability

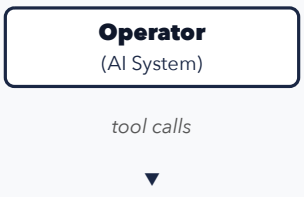
AAL-4 (cryptographic attestation with independent trust root) SHALL require that the notary service be operated by an entity structurally independent of the AI system operator — an Independent Attestation Provider (IAP) per Section 3.14. A single independent notary satisfies AAL-4. Multi-entity notary sets provide additional resilience against compromise but are not required for AAL-4 conformance. Single-IAP AAL-4 therefore establishes attestation independence, not attestation resilience.

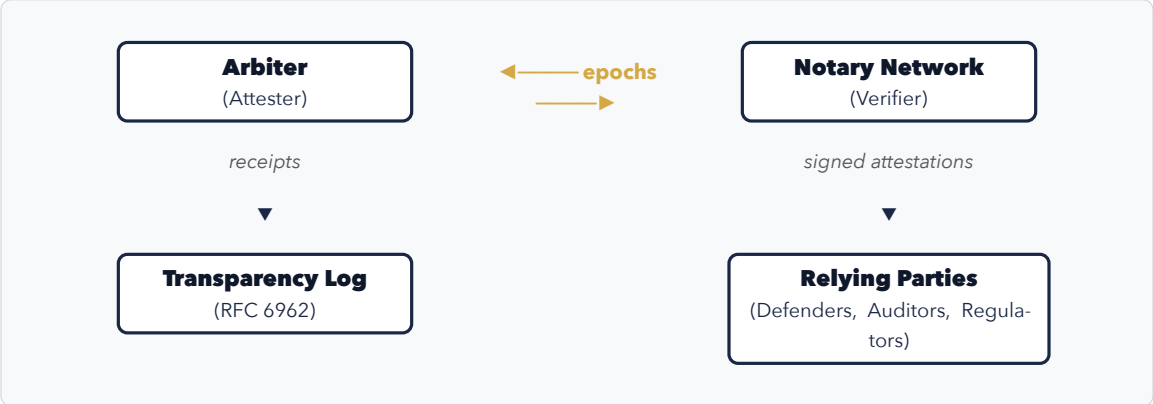
Deployments MAY graduate from AAL-3 to AAL-4 by engaging an independent notary service as specified in ATT-5. The transition SHALL be attested in the transparency log with notary set attestations from both the pre-transition and post-transition configurations.

> ***Note:*** *AAL-1 through AAL-4 describe technical verifiability tiers. They do not correspond to legal burdens of proof, standards of admissibility, or regulatory compliance determinations. Whether an AAL-4 attestation artifact satisfies a particular legal or regulatory standard is a question of applicable law.*

4.2 The Attestation Model

OVERT adopts and extends the IETF RATS (RFC 9334) architecture:





Arbiter (Attester). Deployed at the operator's trust boundary. Intercepts AI system actions, evaluates them against policy, and generates attestation envelopes. The Arbiter sees plaintext — it operates within the operator's security perimeter, analogous to a firewall or security proxy.

Notary Network (Verifier). Structurally independent of the operator. Operated by an Independent Attestation Provider (IAP). Validates attestations using t-of-n notary verification as specified in the registered Protocol Profile. Derives the Arbiter's binary identity independently — the Arbiter cannot self-attest. Publishes epoch nonces and digest ledgers for auditor verification.

Transparency Log. An append-only Merkle tree (RFC 6962) of signed receipts. Provides inclusion proofs (receipt exists in log), consistency proofs (log was not tampered with between time points), and split-view detection.

Relying Parties. Defenders, incident responders, auditors, regulators, procurement teams, insurers, and other parties that need to verify AI control-execution claims without trusting the operator or accessing protected content.

The specific cryptographic constructions, envelope schemas, and protocol details implementing this architecture are specified in registered OVERT Protocol Profiles. Conformant implementations SHALL use a registered OVERT Protocol Profile. Protocol Profile 1.0 is the initial registered profile (see Annex B).

4.3 Trust Architecture

Component	Trust Requirement	Rationale
Arbiter	Operator trusts their own deployment	Same trust model as enterprise firewall
Notary Network	Independent third-party trust; multi-party (t-of-n) where deployed	No single notary can forge attestations; structural independence from operator required for AAL-4

Component	Trust Requirement	Rationale
AI Model/Provider	Untrusted (System Under Test)	The entity being governed is the System Under Test; the attestation system does not trust its self-reports
Transparency Log	Public verifiability	Anyone can audit log consistency

The "untrusted SUT" designation applies to the relationship between the attestation layer and the AI model/provider. The attestation system does not trust the model's self-reports, the provider's claims, or the operator's logs. It produces independent verifiable records.

> ***Note:*** *The "untrusted SUT" designation is specific to the OVERT attestation relationship and is distinct from the NIST SP 800-207 Zero Trust Architecture for network security. SP 800-207 addresses network access assumptions; OVERT addresses attestation independence assumptions. The two are complementary but operate at different layers.*

4.4 Deployment Topology

Mode 1: Sidecar. For self-hosted models. The Arbiter runs as an enforcement module adjacent to the model runtime within the operator's infrastructure. Tool calls are intercepted at the service boundary.

Mode 2: Gateway. For SaaS-based models (OpenAI, Anthropic, Google). The Arbiter operates as a forward proxy. The operator routes orchestration traffic through the gateway, which governs tool execution even when the model runs in a third-party environment. Mode 2 may also be used for self-hosted models where the operator prefers a proxy deployment over a sidecar deployment. The distinction is architectural topology, not hosting model.

Both modes produce identical attestation receipts. The attestation artifacts concern what the operator's system did — not about the model's internals.

4.5 Threat Model and Trust Assumptions

OVERT assumes the following threat model. Conformant implementations SHALL address each threat vector through the specified mitigation. Where a mitigation is marked SHOULD in the normative body (e.g., reproducible builds, binary transparency logs), the threat is addressed through disclosure and compensating controls rather than a hard requirement. The "Required Mitigation" column describes the intended mitigation approach; the normative strength (SHALL, SHOULD, MAY) of each specific control is defined in the referenced section.

Threat Vector	Description	Required Mitigation
Arbiter compromise	Malicious operator modifies or replaces arbiter binary	Notary-derived binary identity via hardware-rooted or hypervisor-attested measurement (NOT client-supplied claims)
Epoch-nonce prediction	Operator predicts sampling nonce to game which requests are evaluated	CSPRNG generation + commitment-reveal scheme (nonce committed at epoch start, revealed after close)
Co-epoch forgery	Attacker fabricates attestation receipts for a prior epoch	Strict current-epoch rule with bounded skew; stale submissions rejected
PRF gaming	Operator manipulates request ordering/content to avoid sampling	Policy-scoped key derivation as specified in the Protocol Profile; PRF deterministic from request commitment
Notary collusion	Subset of notaries collude to forge or suppress attestations	t-of-n notary agreement requirement; no single entity controls t nodes
Transparency log manipulation	Log operator tampers with append-only log	Split-view detection via published Signed Tree Heads; independent monitors
Clock manipulation	Operator skews system clock to place events in wrong epochs	Notary-issued epoch tokens with independent timestamp; bounded skew tolerance
Key compromise	Operator's KMS keys are exfiltrated	Key rotation procedures; epoch-scoped key derivation limits blast radius
Replay/rollback	Attacker replays old valid attestations	Epoch binding prevents cross-epoch replay; receipt includes monotonic sequence
DPL omission	Operator omits requests from Digest Publication Ledger	Coverage ratio computation; gap detection by auditors
Notary censorship	Notary selectively refuses to sign valid attestations	t-of-n requirement prevents single-notary censorship; uptime metrics
IAP compromise / coercion / acquisition	Compromised, coerced, acquired, or negligent IAP issues fraudulent receipts or suppresses anomaly evidence	IAP compromise response plan (Section 4.7.1); multi-IAP option for higher assurance; receipt quarantine for affected epochs; annual transparency reports

Threat Vector	Description	Required Mitigation
Transparency log equivocation	Log operator presents different views (different STHs or inclusion proofs) to different parties	Mandatory independent log monitors (min. 2 for AAL-4); STH gossip protocol; consistency verification publication (Section 4.7.2)
Arbiter side-channel / memory scrape	Attacker exploits arbiter runtime to exfiltrate plaintext content or extract tenant_pepper key material	Process isolation and memory protection; attested key injection channel; TEE (SHOULD for AAL-3/4); runtime integrity monitoring (Section 4.7.3)
Build pipeline compromise	Compromised CI/CD injects malicious arbiter binaries that bypass enforcement	Reproducible builds (SHOULD); binary transparency logs (SHOULD); provenance verification before deployment
Prompt-injection-induced tool abuse	Untrusted input induces the agent to invoke tools, destinations, or data flows that are syntactically valid but unauthorized for the requesting context	Input filtering, pre-execution policy enforcement, parameter validation, provenance-aware authorization, and architectural separation (PRO-4, TOOL-1, TOOL-2, CAP-1, CAP-2)
Delegated-capability abuse	An agent relays, inherits, or composes capabilities beyond those originally granted through delegation, spawning, or topology changes	Capability mediation, spawn authorization, agent topology attestation, and inter-agent trust boundaries (CAP-1, CAP-2, MULTI-1, MULTI-2, DRIFT-3.4)
Approval-path abuse	A sensitive action is pushed through a weak or fatigued human approval path, including rubber-stamping or misbound reviewer identity	Approval gates, reviewer identity binding, approval velocity controls, review-quality monitoring, and separation of duties (TOOL-4, HITL-2, HITL-4, DRIFT-5)
Mediation scope evasion (selection bias)	Operator narrows mediation scope to exclude unfavorable traffic, making signals appear cleaner	Scope statement published to transparency log; scope changes attested with justification; coverage ratio references independent ingress metrics (Section 4.7.4)
Coverage blind spots / denominator ambiguity	The implementation cannot independently demonstrate what traffic or action volume formed the denominator for coverage and measurement claims	Published mediation scope statement, denominator source declaration, independent ingress metrics or profile-defined equivalent, and explicit disclosure of unverifiable

Threat Vector	Description	Required Mitigation
		denominators (Sections 4.7.4, 19.7.4, 22.1)

Trust assumptions:

Conformant implementations SHALL anchor arbiter and configuration measurements in an independently verifiable root of trust. The measurement pipeline SHALL satisfy the properties defined in Section 18.2: not controlled by the attester, rooted in a hardware or cryptographic trust anchor, and reproducible by an independent auditor. Client-supplied identity claims alone are insufficient for AAL-4 conformance.

> ***Note:*** *See Section 18.2 for examples of acceptable measurement pipelines including hardware-rooted attestation, hypervisor-attested measurement, and equivalent infrastructure defined in a registered Protocol Profile.*

4.6 Risk Signal Architecture

OVERT is designed to produce quantitative runtime signals from the attestation stream within the declared mediation scope. Whether a given signal is independently verifiable depends on the denominator source: signals whose denominators are independently verifiable (e.g., derived from independent ingress metrics or notary-observed counts) are classified as **independently verifiable signals**; signals whose denominators are operator-declared only are classified as **operator-dependent signals**. Both classes are useful; the distinction determines what a relying party can verify without trusting the operator.

Conformant implementations SHALL produce risk signals satisfying the following properties:

1. **Content-free derivation.** All signals SHALL be derivable without access to the operator's protected content. Signals are computed from the transparency log, published epoch data, mediation scope statements, and the registered Protocol Profile.
2. **Verifiability classification.** Each signal SHALL be classified as independently verifiable or operator-dependent based on the denominator source. A signal whose denominator is operator-declared only SHALL NOT be presented as independently verifiable in conformance documentation or public claims.
3. **Temporal granularity.** Signals SHALL be expressible as time series at epoch-level granularity.
4. **Statistical rigor.** Signals derived from sampling SHALL carry exact confidence intervals (not approximate), sample sizes, and auditor-reproducible methodology.

5. **Scope binding.** All signals SHALL reference the mediation scope statement, which defines signal denominators, and SHALL disclose the denominator source classification.

Risk signals support governance monitoring, security operations, audit, regulatory reporting, and external risk analysis. Signal definitions, formulas, and derivation procedures are specified in the registered Protocol Profile or companion signal specification. See Annex D for the signal framework and design rationale.

Level 3 and Level 4 conformance SHALL produce, at minimum, the following mandatory signal set per epoch:

1. **Coverage ratio** — the ratio of attested actions to total in-scope actions, referencing the declared denominator source and its verifiability classification (independently verifiable or operator-dependent).
2. **Violation rate with confidence interval** — the estimated policy violation rate with exact confidence bounds (per MEA-2.4).
3. **Gap accounting** — the count and percentage of attestation gap events (per ATT-3.4).
4. **Optimistic enforcement ratio** — the percentage of in-scope actions processed under optimistic enforcement (where applicable; per ATT-3.5(d)), reported both as claim-period average and worst single-epoch value.
5. **Exposure-window duration** — total duration and percentage of the claim period during which attestation coverage lapsed (fail-open periods, IAP unavailability, or unattested operation).

Registered Protocol Profiles MAY define additional signals. A Protocol Profile that does not produce the mandatory signal set is non-conformant for Level 3 and Level 4 claims.

Interpretation of signals for contractual, legal, regulatory, or financial purposes remains external to this standard. OVERT defines the signal architecture; it does not prescribe operational, legal, or actuarial conclusions.

4.7 Security Considerations

This section defines the minimum operational security baseline for OVERT deployments. These controls address threats identified in Section 4.5 that are not resolved by cryptographic format requirements alone. They establish the baseline protections needed for the OVERT trust model itself to remain credible, including response to IAP compromise, transparency-log monitoring, arbiter hardening, mediation-scope attestability, and anomaly triage. All requirements in this section are normative.

4.7.1 IAP Compromise Response

Operators SHALL maintain an IAP compromise response plan. The plan SHALL define, at minimum:

(a) Criteria for initiating a compromise response, including but not limited to: confirmed key compromise, suspected coercion, change-of-control event affecting the IAP, and notification from the IAP of a suspected compromise.

(b) Quarantine procedures for receipts issued during the suspected compromise period. Receipts issued during a suspected compromise period SHALL be quarantined and SHALL NOT be presented as evidence of conformance pending investigation and disposition.

(c) Notification procedures for downstream relying parties that have consumed receipts from the affected IAP during the compromise window.

(d) Re-attestation procedures for epochs affected by the compromise, using an unaffected IAP or through independent verification.

(e) Criteria for restoring trust in a previously compromised IAP, or for permanently revoking trust and transitioning to an alternative IAP.

IAPs SHALL notify affected operators within 72 hours of detecting or suspecting a compromise event. The notification SHALL include the earliest and latest times bounding the suspected compromise window, the nature of the suspected compromise, and a list of affected operator identities or a statement that all operators should be considered potentially affected.

Operators that rely on a single IAP SHALL document the residual risk of single-IAP dependence in their conformance declaration and SHALL satisfy the following resilience requirements:

(f) **Portability escrow.** The operator SHALL maintain a tested portability package (key material escrow, configuration artifacts, and transparency log export) sufficient to onboard a replacement IAP without loss of historical attestation data.

(g) **Migration rehearsal.** The operator SHALL conduct an IAP migration rehearsal at intervals not exceeding 12 months and SHALL attest the rehearsal execution and measured activation time. The rehearsal SHALL demonstrate that the portability escrow enables functional attestation under a replacement IAP.

(h) **Failover procedure.** The operator SHALL define an IAP failover procedure with a target activation time documented in the conformance declaration. The target activation time SHALL be informed by the operator's measured rehearsal results and the current availability of qualified replacement IAPs, not by a fixed calendar period. If no qualified replacement IAP is available at the time of conformance, the operator SHALL disclose this limitation.

During the failover period, the deployment operates under fail-open or fail-closed procedures (RES-5) and SHALL report the unattested duration as an exposure window. Such deployments satisfy

AAL-4 for attestation independence, not for attestation resilience. Level 4 conformance claims based on single-IAP deployments SHALL disclose the IAP topology (single-IAP vs. multi-IAP), the most recent rehearsal date and measured activation time, and any period during the claim window in which no qualified replacement IAP was available.

4.7.2 Transparency Log Monitor Diversity

AAL-4 deployments SHALL engage at minimum two independent transparency log monitors. For the purposes of this requirement, "independent" means that the monitors: (a) are operated by distinct legal entities with no common controlling interest, (b) do not share signing key infrastructure, and (c) operate from network vantage points not co-located with the transparency log operator's primary infrastructure.

Monitors SHALL perform the following verification functions:

1. **Consistency verification.** Monitors SHALL verify that each Signed Tree Head (STH) is consistent with all previously observed STHs for the same log, at intervals not exceeding the epoch boundary frequency.
2. **Inclusion verification.** Monitors SHALL periodically verify that receipts known to have been submitted to the log are included in the published tree. The sampling rate for inclusion verification SHALL be documented.
3. **Cross-monitor gossip.** Monitors SHALL exchange observed STHs with at least one other independent monitor. Detection of an STH discrepancy SHALL be treated as a log equivocation event and SHALL trigger immediate notification to all affected operators.

Monitors SHALL publish consistency verification results at a location accessible to relying parties. AAL-3 deployments SHOULD engage at least one independent transparency log monitor.

4.7.3 Arbiter Hardening

Arbiter deployments SHALL implement process isolation and memory protection appropriate to the sensitivity of the content processed. At minimum:

(a) The arbiter process SHALL execute in an isolated process boundary with restricted system call access. The arbiter SHALL NOT share a process address space with application code.

(b) Memory regions containing operator key material (tenant_pepper, content-binding keys) SHALL be protected from access by other processes.

(c) Operator key material SHALL be injected into the arbiter via an attested channel — one in which the recipient can be cryptographically verified to be running the expected binary in the expected isolation state before key material is transmitted. Acceptable mechanisms include hardware-attested

sealed channels, mutually authenticated TLS with identity bound to co-epoch state, or KMS with policy-gated release tied to arbiter binary hash.

(d) Operator key material SHALL NOT be passed via environment variables in production deployments, persisted to disk in plaintext, logged at any verbosity level, or included in core dumps or crash reports.

(e) The arbiter SHALL zeroize sensitive key material from memory upon epoch rotation and upon process termination.

(f) For AAL-4 deployments, the arbiter SHALL either execute within a hardware-attested trusted execution environment (TEE) or the conformance claim SHALL explicitly disclose that arbiter isolation is software-only and not hardware-rooted. For AAL-3 deployments, the arbiter SHOULD be executed within a hardware-attested trusted execution environment (TEE).

4.7.4 Mediation Scope Attestability

The mediation scope statement (as defined in Section 3.16) SHALL be published to the transparency log. The published scope statement SHALL include a machine-readable definition of the traffic classes within scope, any exclusions with stated justification, and the effective date.

Changes to the mediation scope SHALL be attested by the operator and logged to the transparency log with the previous scope statement hash, the new scope statement hash, a machine-readable justification, and the effective date of the change. Relying parties SHALL have access to the complete mediation scope history.

All Level 3 and Level 4 conformance claims SHALL identify the mediation scope statement hash, the declared coverage percentage of the mediation scope relative to its denominator, the denominator source used for coverage and measurement claims, and whether that denominator source is independently verifiable or operator-declared only. Where independently verifiable ingress metrics are available (e.g., load balancer request counts, API gateway telemetry), the coverage ratio SHALL reference those metrics. Where independent ingress metrics are not available, the attestation SHALL disclose this limitation. Level 4 claims SHALL use independently verifiable ingress metrics or a registered-Protocol-Profile equivalent denominator source; absent such evidence, the implementation SHALL NOT claim Level 4 conformance for that scope.

4.7.5 Anomaly Triage Obligation

Operators SHALL establish and maintain documented procedures for triaging, dispositioning, and escalating attested anomalies. Attested anomalies include but are not limited to: policy violations, override patterns exceeding baseline thresholds, drift signals breaching alert thresholds, exposure windows, coverage ratio shortfalls, and receipt verification failures.

The anomaly triage procedure SHALL define:

(a) **Classification criteria.** A severity classification scheme with defined criteria based on type, frequency, and potential impact.

(b) **Response timelines.** Maximum time-to-acknowledge and time-to-disposition for each severity level. Critical anomalies SHALL be acknowledged within 24 hours and dispositioned within 7 days. Security-critical anomalies — including binary identity mismatch, co-epoch binding violation, transparency log equivocation, and arbiter integrity failure — SHALL be acknowledged within 1 hour and SHALL trigger immediate containment action (circuit breaker, scope isolation, or fail-closed) pending disposition.

(c) **Disposition categories.** At minimum: confirmed violation (remediate), false positive (document rationale), accepted risk (document rationale and approval authority), and escalation (to identified authority).

(d) **Escalation paths.** Named roles or functions responsible for escalation decisions at each severity level.

(e) **Record retention.** Triage records SHALL be retained for the period defined in the operator's retention schedule and SHALL be available for audit.

> ***NOTE — Adverse inference implications.*** *An attested anomaly constitutes a record of a condition observed and recorded by the system. Failure to act on attested anomalies — or failure to maintain triage procedures that ensure anomalies are reviewed — may constitute constructive notice of the conditions evidenced by those anomalies. Operators should consult legal counsel regarding the evidentiary implications of attested anomaly records in their jurisdiction. This note is informative and does not create legal obligations beyond those stated normatively in this section.*

4.8 Cross-Boundary Attestation Protocol

Many real-world AI deployments involve multiple trust boundaries in sequence — for example, an ambient scribe producing a clinical note that feeds a clinical decision support system, which queries a drug interaction database, which in turn calls a genomics API. Each boundary operator may independently deploy OVERT attestation. This section defines how attestation receipts are linked across trust boundaries to enable end-to-end verification without requiring protected content to cross any boundary.

4.8.1 Purpose

Cross-boundary attestation enables relying parties to verify that governance controls executed across an entire multi-provider workflow, not merely within a single operator's boundary. The protocol achieves this by linking receipts across trust boundaries using cryptographic references — specifically, by including the upstream receipt's attestation_id hash in the downstream receipt. No protected content crosses any trust boundary; only receipt hashes (attestation_id references) are exchanged.

4.8.2 Parent Attestation Reference

Each receipt generated within a downstream trust boundary MAY reference an upstream receipt by including the upstream receipt's attestation_id hash as a `parent_attestation_id` field in the downstream receipt. The `parent_attestation_id` SHALL be the SHA-256 hash of the upstream receipt's attestation_id as published in the upstream operator's transparency log. Where multiple upstream receipts contributed to a single downstream action, the downstream receipt MAY include multiple `parent_attestation_id` entries.

The `parent_attestation_id` field is OPTIONAL for workflows that do not cross trust boundaries. For cross-boundary workflows at Level 3 or above, the `parent_attestation_id` field SHALL be populated when the downstream operator has access to the upstream receipt's attestation_id.

4.8.3 Cross-Boundary DAG Reconstruction

Relying parties can reconstruct an end-to-end directed acyclic graph (DAG) of attestation across providers by following the `parent_attestation_id` hash chains. Each node in the DAG represents a receipt within a single trust boundary; each edge represents a `parent_attestation_id` reference linking a downstream receipt to an upstream receipt. The DAG enables verification that governance controls executed at every boundary in a multi-provider workflow.

DAG reconstruction SHALL NOT require access to protected content from any boundary. Relying parties reconstruct the DAG using only: (a) receipt metadata and `parent_attestation_id` fields from each boundary's transparency log, (b) publicly verifiable receipt signatures and co-epoch bindings, and (c) published cross-boundary scope statements (Section 4.8.5).

4.8.4 No Content Crossing

Only receipt hashes (attestation_id references) cross trust boundaries under this protocol. The `parent_attestation_id` is a hash of a receipt identifier — it does not contain, encode, or enable reconstruction of any protected content from the upstream boundary. The non-egress property (Section 17, Design Principle 2) is preserved across all trust boundaries in the chain.

4.8.5 Cross-Boundary Scope Statement

Each boundary operator participating in cross-boundary attestation SHALL publish a cross-boundary scope statement to its transparency log. The cross-boundary scope statement SHALL declare:

(a) Which upstream attestation sources the operator accepts and links (identified by upstream operator identity and upstream transparency log URI).

(b) The upstream receipt validation policy: whether the operator validates upstream receipt signatures, co-epoch bindings, and transparency log inclusion proofs before linking, or accepts upstream attestation_id references without independent validation.

(c) The effective date and version of the cross-boundary scope statement.

Changes to the cross-boundary scope statement SHALL be attested and logged with the same change-attestation requirements as mediation scope statement changes (Section 4.7.4).

4.8.6 Receipt Chain Validation

Verifiers performing cross-boundary DAG validation SHALL validate the full chain by checking, for each receipt in the DAG:

(a) The receipt's signature validity (per the receipt's boundary operator's notary network).

(b) The receipt's co-epoch binding integrity (binary identity, network state, epoch currency).

(c) The `parent_attestation_id` reference integrity: the referenced upstream receipt EXISTS in the upstream operator's transparency log and the `parent_attestation_id` value matches the SHA-256 hash of the upstream receipt's attestation_id.

(d) The downstream operator's cross-boundary scope statement declares acceptance of the upstream attestation source.

A cross-boundary verification is valid only if every receipt in the chain passes all four checks. Partial chain validation (where some links are verified and others are not) SHALL be reported as partial, not as full cross-boundary verification.

4.8.7 Failure Handling

If an upstream receipt is unavailable or invalid at the time the downstream receipt is generated, the downstream receipt SHALL include a `parent_reference_status` field with one of the following values:

Status	Description
`VALID`	Upstream receipt was available, its signature and co-epoch binding were verified, and the `parent_attestation_id` was successfully computed and included

Status	Description
`UNAVAILABLE`	Upstream receipt was not available within the profile-defined timeout; `parent_attestation_id` could not be populated
`INVALID`	Upstream receipt was available but failed signature verification, co-epoch binding verification, or transparency log inclusion verification
`TIMEOUT`	Upstream transparency log did not respond within the profile-defined timeout for inclusion proof verification

When the `parent_reference_status` is anything other than `VALID`, the downstream receipt SHALL still be generated (attestation of the downstream boundary's own governance controls proceeds regardless of upstream availability), but the cross-boundary chain is incomplete at that link. Relying parties SHALL treat incomplete links as gaps in cross-boundary verification, not as failures of the downstream boundary's own attestation.

4.8.8 Normative Requirements

All cross-boundary attestation controls in this section are normative at AAL-4 for Level 3 and Level 4 conformance claims involving cross-boundary workflows. Level 1 and Level 2 claims are not required to implement cross-boundary attestation. Implementations that do not participate in cross-boundary workflows are not required to implement this section.

The specific `parent_attestation_id` field encoding, `parent_reference_status` enumeration, cross-boundary scope statement schema, and DAG reconstruction procedures are specified in the registered Protocol Profile.

OVERT is an open standard for public adoption and co-development. Implementation details are specified in registered OVERT Protocol Profiles.

Editorial contact: overt-review@glacis.io Protocol Profile Registry: See Annex B and Section 22.6

Version 0.7.0 — March 2026 — Public Review Draft

PART 2: GOVERNANCE DOMAINS

OVERT organizes its core requirements into governance domains that together define the organizational and infrastructure control plane for verifiable AI governance and AI runtime defense. Part 2 covers organizational governance, system identification, boundary enforcement, attestation generation and verification, measurement, and response. Read together, these domains define how an operator declares policy, constrains AI actions at the boundary, measures in-scope behavior, and preserves evidence of control execution.

5. Domain 1: GOVERN – Organizational Governance

Scope: Policies, accountability structures, training, culture, and supply chain governance. Maps to NIST AI RMF GOVERN and ISO 42001 Clauses 4–7.

These controls are organizational in nature. Attestation assurance level requirements are AAL-1–AAL-2 for policy and process controls, with AAL-4 required where machine-verifiable artifacts are possible.

GOV-1: AI Governance Policy

Requirement: The organization SHALL establish, document, and maintain an AI governance policy covering all AI systems within scope.

Attestation Assurance Level: AAL-1 (policy document) + AAL-4 (machine-readable policy artifact published to transparency log)

ID	Control	Attestation Artifact	Level
GOV-1.1	Publish AI governance policy covering intended uses, risk tolerances, accountability structures, and applicable regulations	Policy document in human-readable format	AAL-1
GOV-1.2	Publish machine-readable policy artifact (OSCAL or OVERT policy schema) to transparency log with cryptographic timestamp	Signed policy artifact with transparency log inclusion proof	AAL-4

ID	Control	Attestation Artifact	Level
GOV-1.3	Review and update policy at planned intervals (minimum: annually) with documented change justification	Transparency log entries showing versioned policy updates	AAL-4

NIST AI RMF: GOVERN 1.1, 1.2, 1.4 | **ISO 42001:** 5.2, A.2.2, A.2.4

GOV-2: Accountability and Roles

Requirement: The organization SHALL assign and document roles and responsibilities for AI risk management, including a designated accountable individual for each AI system in scope.

Attestation Assurance Level: AAL-2

ID	Control	Attestation Artifact	Level
GOV-2.1	Assign accountable owner for each AI system with documented authority and responsibility	Organizational chart or RACI matrix	AAL-2
GOV-2.2	Define and document change approval authority — which system changes require formal review and by whom	Change approval policy with designated approvers per change type	AAL-2
GOV-2.3	Ensure separation of duties: personnel responsible for AI development SHALL NOT serve as sole approver of their own work	Documented approval records showing independent sign-off	AAL-2

NIST AI RMF: GOVERN 2.1, 2.3 | **ISO 42001:** 5.3, A.3.2 | **AIUC-1:** E004

GOV-3: Risk Taxonomy

Requirement: The organization SHALL establish and maintain a risk taxonomy categorizing AI-specific risks with severity levels, examples, and remediation procedures.

Attestation Assurance Level: AAL-2 + AAL-4 (taxonomy published as machine-readable artifact)

ID	Control	Attestation Artifact	Level
GOV-3.1	Define risk categories covering: harmful outputs, out-of-scope outputs, hallucinated outputs, unauthorized tool actions, data leakage, bias, and domain-specific risks	Risk taxonomy document	AAL-2
GOV-3.2	Assign severity levels to each risk category with escalation criteria	Severity matrix with escalation procedures	AAL-2

ID	Control	Attestation Artifact	Level
GOV-3.3	Publish machine-readable risk taxonomy to transparency log and reference it in attestation policy configuration	Signed taxonomy artifact with log inclusion proof	AAL-4
GOV-3.4	Review taxonomy at intervals defined in the operator's risk management policy, not to exceed 12 months or as defined by applicable regulation; update based on incidents, emerging threats, and regulatory changes	Transparency log entries showing versioned taxonomy updates	AAL-4
GOV-3.5	Require that all attestation policy artifacts (the policy_hash referenced in enforcement receipts) be signed by a designated Qualified Risk Officer and reference a published safety baseline (e.g., NIST AI Safety Profile, OWASP Agentic Top 10, or sector-specific baseline). The receipt validates enforcement; the signature validates the rule set. [See Annex C: Design Rationale, "Policy-Quality Gap" analysis]	Policy artifact signed by named risk officer with baseline reference in transparency log	AAL-4

NIST AI RMF: MAP 1.1, 5.1 | **ISO 42001:** 6.1.2, A.5.2 | **AIUC-1:** C001

GOV-4: Supply Chain and Third-Party Governance

Requirement: The organization SHALL establish governance processes for third-party AI components, including foundation models, data sources, and tools.

Attestation Assurance Level: AAL-2

ID	Control	Attestation Artifact	Level
GOV-4.1	Conduct documented due diligence on foundation model providers covering: data handling, security practices, safety testing, and contractual commitments	Vendor assessment records	AAL-2
GOV-4.2	Maintain inventory of all third-party AI components with version tracking and provenance documentation	Component inventory with version history	AAL-2
GOV-4.3	Establish contractual requirements for third-party components including: notification of material changes, incident disclosure, and cooperation with audits	Contract excerpts or attestation from legal review	AAL-2

NIST AI RMF: GOVERN 6.1, 6.2, MAP 4.1 | **ISO 42001:** A.10.2, A.10.3 | **AIUC-1:** E006

GOV-5: AI Disclosure

Requirement: The organization SHALL implement disclosure mechanisms informing users when they interact with AI systems.

Attestation Assurance Level: AAL-2 (product demonstrations)

ID	Control	Attestation Artifact	Level
GOV-5.1	Implement disclosure for text-based AI interactions ("You are chatting with an AI")	Product screenshot or recording	AAL-2
GOV-5.2	Implement disclosure for voice-based AI interactions (spoken notification at session start)	Audio recording or transcript	AAL-2
GOV-5.3	Label AI-generated content in machine-readable format (C2PA Content Credentials, metadata, or watermarks)	Content sample with embedded metadata	AAL-2
GOV-5.4	Disclose when autonomous AI agents perform actions without step-by-step human oversight	Product demonstration showing agent disclosure	AAL-2
GOV-5.5	Respond accurately when users ask "Are you AI?"	Product demonstration	AAL-2
GOV-5.6	Include a receipt reference (receipt_id or receipt_hash) in AI response metadata, enabling end users to dispute a specific interaction by citing its cryptographic identifier. The operator can then locate the exact attestation record, verifiable record, and policy evaluation for that transaction	Receipt reference in response metadata; dispute resolution procedure	AAL-4

NIST AI RMF: GOVERN 4.2 | **ISO 42001:** A.8.2 | **AIUC-1:** E016

6. Domain 2: IDENTIFY – Risk Identification and Mapping

Scope: Context establishment, AI system categorization, impact assessment, and risk mapping. Maps to NIST AI RMF MAP and ISO 42001 Clause 6.

IDE-1: System Context and Categorization

Requirement: The organization SHALL document the intended purpose, deployment context, capabilities, and limitations of each AI system in scope.

ID	Control	Attestation Artifact	Level
IDE-1.1	Document intended purposes, target users, deployment settings, and applicable laws/regulations	System context document	AAL-1
IDE-1.2	Categorize system capabilities: text generation, voice generation, image generation, automation/agentic, or multimodal	Capability classification in machine-readable format	AAL-2
IDE-1.3	Document system knowledge limits and conditions under which outputs may be unreliable	Technical limitations document	AAL-1

NIST AI RMF: MAP 1.1, 2.1, 2.2 | **ISO 42001:** 4.1, A.6.2.2

IDE-2: AI System Impact Assessment

Requirement: The organization SHALL assess and document potential consequences of each AI system on individuals, groups, and society.

ID	Control	Attestation Artifact	Level
IDE-2.1	Conduct impact assessment for each AI system covering: potential benefits, potential harms, affected populations, and severity of adverse outcomes	Impact assessment document	AAL-2
IDE-2.2	Consider domain-specific and jurisdictional requirements in impact assessments	Jurisdictional analysis	AAL-2
IDE-2.3	Incorporate impact assessment results into risk treatment planning	Risk treatment plan referencing impact assessment	AAL-2

NIST AI RMF: MAP 5.1, 5.2 | **ISO 42001:** 6.1.4, A.5.2–A.5.5

7. Domain 3: PROTECT – Boundary Enforcement and Containment

Scope: Runtime enforcement of governance policy at the boundary between AI systems and external resources. This is the domain where OVERT departs from existing frameworks by requiring enforcement infrastructure, not just policy documentation.

All controls in this domain require AAL-4 attestation artifacts.

PRO-1: Boundary Enforcement

Requirement: All AI system interactions with external resources (tool calls, API requests, data access, network egress) SHALL pass through an enforcement layer that evaluates actions against defined policy before execution.

ID	Control	Attestation Artifact	Level
PRO-1.1	Deploy an enforcement arbiter at the boundary between AI system and external resources	Co-epoch attested binary hash proving arbiter deployment	AAL-4
PRO-1.2	Evaluate every outbound action against customer-defined policy before execution	Per-action attestation receipt (permit or deny)	AAL-4
PRO-1.3	Block actions that violate policy; generate denial receipt with policy reference	Denial receipts in transparency log	AAL-4
PRO-1.4	Generate permit receipt for allowed actions, cryptographically bound to policy version and system configuration	Permit receipts with co-epoch binding	AAL-4

NIST AI RMF: MANAGE 2.4 | **ISO 42001:** 8.1 | **AIUC-1:** B006, D003

PRO-2: Network Isolation and Egress Control

Requirement: AI system network egress SHALL be restricted to approved destinations and attested at each epoch.

ID	Control	Attestation Artifact	Level
PRO-2.1	Implement destination allowlists restricting AI system network egress to approved endpoints	Attested network policy hash (NETATT)	AAL-4
PRO-2.2	Attest network isolation state at each epoch covering: egress policy, network policy definitions (hash the policy controller input, not	Co-epoch NETATT with multi-layer hash	AAL-4

ID	Control	Attestation Artifact	Level
	dynamic ephemeral rules), network controller identity, eBPF programs, CNI configuration, runtime environment variables affecting AI behavior, and TLS certificate pins		
PRO-2.3	Detect and attest any network configuration changes within an epoch	Configuration drift detection via NETATT hash comparison	AAL-4

PRO-3: Rate Limiting and Velocity Controls

Requirement: AI system actions SHALL be subject to rate limits and velocity controls with attested enforcement.

ID	Control	Attestation Artifact	Level
PRO-3.1	Enforce per-action, per-user, and per-epoch rate limits on tool calls and API requests	Rate limit enforcement receipts	AAL-4
PRO-3.2	Implement escalating restrictions for anomalous velocity patterns	Velocity enforcement attestations	AAL-4
PRO-3.3	Require human approval gates for actions exceeding defined thresholds	Approval gate attestations with identity binding	AAL-4

AIUC-1: B004, D003.2

PRO-4: Input and Output Filtering

Requirement: AI system inputs and outputs SHALL be filtered for safety policy violations with attested enforcement.

ID	Control	Attestation Artifact	Level
PRO-4.1	Filter inputs for adversarial content, prompt injection, and policy violations before model processing	Filter enforcement receipts	AAL-4
PRO-4.2	Filter outputs for harmful content, out-of-scope content, PII leakage, and policy violations before delivery	Filter enforcement receipts	AAL-4
PRO-4.3	Sanitize outputs to prevent security vulnerabilities (XSS, injection, unsafe URLs) in downstream systems	Sanitization enforcement receipts	AAL-4

AIUC-1: B005, C003, C004, C005, C006, A006

PRO-5: Data Isolation

Requirement: Customer data SHALL be isolated with attested enforcement of tenant boundaries.

ID	Control	Attestation Artifact	Level
PRO-5.1	Enforce logical and/or physical separation of customer data across tenants	Data isolation attestation	AAL-4
PRO-5.2	Attest that AI system prompts and responses do not cross tenant boundaries	Cross-tenant isolation receipts	AAL-4
PRO-5.3	Implement PII detection and filtering with attested enforcement	PII detection receipts (no content egress)	AAL-4

AIUC-1: A005, A006

8. Domain 4: ATTEST – Attestation Generation and Verification

Scope: The core attestation infrastructure. This domain specifies how attestation artifacts are generated, stored, attested, and made verifiable by third parties.

ATT-1: Non-Egress Attestation Architecture

Requirement: The attestation protocol SHALL NOT require transmission of protected content outside the operator environment. Conformant receipt-service interfaces SHALL accept only cryptographic commitments and profile-defined metadata.

ID	Control	Attestation Artifact	Level
ATT-1.1	Canonicalize AI request/response payloads using deterministic encoding as specified in the registered Protocol Profile	Documented encoder specification with version-pinned encoder_id	AAL-4
ATT-1.2	Compute request digests as cryptographic hashes of canonical encodings; derive keyed commitments using a keyed cryptographic function with tenant-scoped keys held exclusively in the operator's KMS. Only keyed commitments cross the trust boundary — never raw digests. This prevents rainbow table reversal of low-entropy content (PII, SSNs) by any party with ledger access	Receipt service schema accepts only keyed commitments; raw digests rejected; `additionalProperties: false`	AAL-4
ATT-1.3	Store attestation artifacts (full payloads, policy evaluations, metadata) in content-addressable storage within the operator's environment	Local CAS deployment with retention policy	AAL-4

ID	Control	Attestation Artifact	Level
ATT-1.4	Constrain attestation egress to a single receipt service endpoint over TLS with certificate pinning as defined in the registered Protocol Profile	Attested certificate pin set in NETATT	AAL-4

> ***Note:*** *For streaming outputs (Server-Sent Events), implementations MAY use rolling commitment constructions or chunked attestation as defined in the registered Protocol Profile. The full-payload commitment model described here is the normative baseline; streaming extensions are profile-specific.*

> ***Note:*** *The keyed commitment requirement (ATT-1.2) specifies properties, not constructions. The keyed function SHALL be computationally infeasible to invert without knowledge of the operator secret. Protocol Profile 1.0 satisfies this requirement using HMAC-SHA256 with keys derived via HKDF. Alternative profiles MAY use different keyed commitment schemes provided they satisfy the non-egress and irreversibility properties defined above.*

ATT-2: Co-Epoch Binding

Requirement: Every attestation receipt SHALL be cryptographically bound to the system's binary identity and network isolation state during a bounded time interval.

ID	Control	Attestation Artifact	Level
ATT-2.1	Establish heartbeat epochs with configurable epoch duration (recommended: 300 seconds) with notary-issued bearer tokens	Epoch heartbeat receipts with notary signatures	AAL-4
ATT-2.2	Arbiter binary identity SHALL be derived by the notary through a measurement pipeline that is (a) not controlled by the attester, (b) rooted in a hardware or cryptographic trust anchor, and (c) reproducible by an independent auditor given the measurement policy. Client-supplied identity claims are insufficient for AAL-4 conformance. See Section 18.2 for acceptable measurement pipelines	Notary-derived binary identity in receipt	AAL-4
ATT-2.3	Bind every receipt to the current epoch, binary identity, and network attestation hash	Co-epoch receipt schema with all three bindings	AAL-4

ID	Control	Attestation Artifact	Level
ATT-2.4	Reject any attestation submission not in the current epoch (strict current-epoch rule; bounded skew tolerance as defined in the registered Protocol Profile, recommended: <=2 seconds)	Deterministic rejection: ERR_STALE_EPOCH	AAL-4

ATT-3: Three-Phase Attestation

Requirement: The attestation system SHALL support synchronous enforcement, synchronous provisional receipts, and asynchronous full attestation to meet latency requirements without compromising attestation artifact quality.

ID	Control	Attestation Artifact	Level
ATT-3.1	**Phase 1 — Enforcement:** Evaluate action against policy synchronously. [Informative targets: <5ms P50 local, <25ms P50 distributed. Specific latency requirements are defined in the registered Protocol Profile.]	Enforcement decision recorded in arbiter	AAL-4
ATT-3.2	**Phase 2 — Provisional Receipt:** Generate locally-signed attestation commitment synchronously with explicit provisional status	Provisional receipt with arbiter signature	AAL-4
ATT-3.3	**Phase 3 — Full Attestation:** Notary network validates and counter-signs asynchronously using t-of-n notary verification with cryptographic constructions specified in the registered Protocol Profile. Implementations SHALL support cryptographic agility including post-quantum migration paths. After January 1, 2031, pure classical signature schemes are non-conformant; hybrid classical + post-quantum constructions as specified in the registered Protocol Profile SHALL be used	Full receipt with t-of-n notary signature and transparency log inclusion proof	AAL-4
ATT-3.4	Track and report provisional receipts that are not upgraded to full attestation within the SLA window as explicit "attestation gap" events	Gap accounting in audit reports	AAL-4

ID	Control	Attestation Artifact	Level
ATT-3.5	**Optimistic Enforcement Mode:** For latency-critical deployments (real-time voice agents, sub-100ms SLA), the system MAY proceed on Phase 2 (Provisional Receipt) without blocking on Phase 3 (Notary), subject to strict constraints: Optimistic enforcement is permitted only for actions declared as idempotent and side-effect-free in the operator's policy configuration. Misclassification of a side-effecting action as optimistic-eligible is a governance failure. **(a) Read Classification:** "Read-only" classification SHALL account for reads that may exfiltrate sensitive data, enumerate systems, or trigger irreversible downstream consequences. A read that returns PII, classified data, or credentials is not safely optimistic. Optimistic enforcement SHALL NOT be applied to tool calls classified as Write, Transact, Delete, or Modify — operations with external side effects SHALL always require synchronous Phase 3 Notary Attestation before execution. **(b) Circuit Breaking:** If Phase 3 subsequently rejects a provisional receipt (notary detects drift, binary mismatch, or policy violation), trigger a circuit breaker (TOOL-3.3) halting subsequent requests from the same agent/session. **(c) Gap Classification:** If Phase 3 does not complete within the profile-defined SLA window, the action SHALL be classified as an attestation gap event under ATT-3.4 and SHALL NOT count as fully attested coverage. **(d) Reporting:** Provisional-only periods SHALL be reported as a distinct coverage class in risk signals, separated from fully attested periods. For Level 3 and Level 4 conformance claims, the percentage of in-scope actions processed under optimistic enforcement during the claimed period SHALL be disclosed in the conformance statement, including both the claim-period average and the worst single-epoch	Optimistic mode declaration in policy with explicit tool-call classification; circuit breaker on notary rejection	AAL-4

ID	Control	Attestation Artifact	Level
	value. **(e) Level 4 Cap:** For Level 4 conformance, optimistic enforcement SHALL NOT exceed 25% of in-scope tool calls in any single epoch, and SHALL NOT exceed 15% of in-scope tool calls averaged over the conformance claim period. These caps apply independently per action class (Read, Write, Transact, Delete, Modify, and any operator-defined classes). Deployments exceeding either threshold for any action class SHALL claim Level 3 for the affected scope until the optimistic percentage is reduced. **(f) Eligibility governance:** The operator's classification of actions as optimistic-eligible SHALL be documented in the mediation scope statement and SHALL be subject to review by the IAP or auditor upon request. Misclassification of a side-effecting action as optimistic-eligible is a governance failure that SHALL be reported as a conformance deviation. For Level 3 conformance, optimistic enforcement SHALL NOT exceed 40% of in-scope tool calls in any single epoch. This structure ensures that "Evidence-Grade" conformance reflects predominantly independently verified attestation and that optimistic enforcement cannot be concentrated in high-risk epochs and diluted by low-risk traffic		

ATT-4: Transparency Log

Requirement: All receipts SHALL be recorded in an append-only transparency log providing inclusion proofs, consistency proofs, and split-view detection.

ID	Control	Attestation Artifact	Level
ATT-4.1	Operate an append-only Merkle tree log (RFC 6962) for all attestation receipts	Signed Tree Heads (STH) with root hash, tree size, and timestamp	AAL-4
ATT-4.2	Provide inclusion proofs for any receipt on demand	Merkle inclusion proof path	AAL-4

ID	Control	Attestation Artifact	Level
ATT-4.3	Provide consistency proofs between any two Signed Tree Heads	Merkle consistency proof	AAL-4
ATT-4.4	Publish Signed Tree Heads at regular intervals for independent monitoring and split-view detection	Published STH records	AAL-4

ATT-5: Notary Network Governance

Requirement: The governance, composition, and independence of the notary network SHALL be explicitly defined, documented, and verifiable. The credibility of AAL-4 attestation artifacts depends entirely on the structural independence of the notaries from the operator being attested.

ID	Control	Attestation Artifact	Level
ATT-5.1	Define and publish the notary network governance model. Four models are normative: **(a) Platform-operated:** An Independent Attestation Provider (IAP) operates all notary nodes. Satisfies AAL-4 where the IAP is structurally independent of the AI system operator. **(b) Consortium:** Nodes operated by a combination of operator, insurer, auditor, and IAP — no single entity controls t nodes. Satisfies AAL-4 where at least one consortium member is structurally independent. **(c) Customer-operated:** Full sovereignty for maximum-security environments. Satisfies AAL-3 (the operator controls the notary). **(d) Hardware-enforced (First-Party Enclave):** Cloud provider or operator uses native hardware-attested TEEs (e.g., AWS Nitro Enclaves, Azure Confidential Computing) as the notary, with attestation quotes verifiable by any relying party. Satisfies AAL-3 with enhanced measurement properties; satisfies AAL-4 where the enclave attestation is validated by an independent third party. Any additional governance model MAY be proposed for registration provided it satisfies the independence and publishability requirements defined in this stan-	Governance model documentation with transparency log proof	AAL-4

ID	Control	Attestation Artifact	Level
	dard. Publish governance model to transparency log		
ATT-5.2	Publish the current notary set: node identities, operating entities, geographic distribution. Attest any changes to the notary set with t-of-n notary signatures from both the outgoing and incoming set	Notary set transition attestation in transparency log	AAL-4
ATT-5.3	AAL-4 SHALL require that the notary service be operated by an entity structurally independent of the AI system operator. For deployments using multiple notaries: no single organizational entity SHALL control t or more notary nodes, and no two notary nodes in the same threshold set SHALL share a common ultimate corporate parent, common hosting infrastructure provider, or common jurisdiction of incorporation. For platform-operated models: publish geographic and infrastructure diversity guarantees. For hardware-enforced models: publish TEE attestation quote verification procedures	Notary independence attestation	AAL-4
ATT-5.4	Publish notary network uptime, availability, and attestation latency metrics at regular intervals	Notary health metrics in transparency log	AAL-4

Architectural note: AAL-4 requires structural independence between the notary service and the AI system operator. A single independent third-party notary satisfies this requirement. Multi-entity notary sets (consortium models) provide additional resilience — no single entity compromise can forge attestations — but are not required for AAL-4 conformance. Platform-operated notaries are operationally simpler but introduce a dependency on the IAP's integrity. Customer-operated notaries satisfy AAL-3, not AAL-4. Hardware-enforced models satisfy AAL-3 unless validated by an independent party. The standard does not mandate a single model — it mandates that the chosen model is explicit, published, and auditable, and that the AAL claim matches the achieved independence.

9. Domain 5: MEASURE – Statistical Safety Assessment

Scope: Continuous, quantitative measurement of AI system behavior with cryptographically verifiable sampling. Maps to NIST AI RMF MEASURE but adds the rigor that MEASURE 2.x references but does not specify.

This standard defines a single normative auditor-reproducible sampling and measurement method: the Statistical Safety Signal Protocol (S3P) defined in MEA-2. MEA-1 specifies the deterministic sampling infrastructure that S3P relies upon. Alternative sampling constructions SHALL be specified in a registered Protocol Profile and demonstrated to preserve completeness verification and auditor reconstruction.

MEA-1: Deterministic Sampling Infrastructure

Requirement: All sampling for AI system monitoring SHALL use deterministic, cryptographically verifiable selection — ensuring that the operator cannot selectively monitor favorable interactions. MEA-1 specifies the key derivation and sampling infrastructure that feeds the S3P measurement method (MEA-2).

ID	Control	Attestation Artifact	Level
MEA-1.1	Derive per-policy sampling keys using a key derivation function scoped by policy identifier, as specified in the registered Protocol Profile	Key derivation documented; key fingerprint published	AAL-4
MEA-1.2	Compute pseudorandom function (PRF) tag for each request using a keyed function with domain separation including policy_id and the request commitment produced per ATT-1.2 (the keyed value, not the raw content digest). This ensures an auditor can verify sampling fairness using only the sampling key and published commitments, without requiring access to a key capable of reversing content. Specific PRF construction is defined in the registered Protocol Profile	PRF tags included in attestation envelopes	AAL-4
MEA-1.3	Determine sample membership by comparing PRF tag against threshold: sampled iff the tag value falls within the configured sampling rate boundary	Deterministic threshold computation	AAL-4

ID	Control	Attestation Artifact	Level
MEA-1.4	Publish per-epoch Digest Publication Ledger (DPL) enabling auditors to verify sample completeness via the S3P epoch nonce reveal (MEA-2.5)	DPL with notary signature	AAL-4

MEA-2: Statistical Safety Signal Protocol (S3P)

Requirement: Safety monitoring SHALL produce quantified statistical statements with exact confidence intervals, derived from cryptographically unbiased sampling. S3P is the single normative auditor-reproducible measurement method defined by this standard.

ID	Control	Attestation Artifact	Level
MEA-2.1	Generate secret epoch nonce via CSPRNG; withhold during epoch to prevent gaming	Epoch nonce commitment (cryptographic hash of epoch_nonce) published at epoch start	AAL-4
MEA-2.2	Compute S3P sampling tag using a keyed function with epoch_nonce and the request commitment (per ATT-1.2) as specified in the registered Protocol Profile; sample iff tag falls within the configured sampling boundary	S3P tag computation	AAL-4
MEA-2.3	Conduct full guardrail evaluation on sampled requests; record n_total, n_sampled, n_violations per policy per epoch	Per-epoch S3P attestation	AAL-4
MEA-2.4	Compute exact binomial confidence intervals using conservative statistical methods requiring no distributional assumptions (e.g., Clopper-Pearson method). Specific formulas and minimum credibility thresholds are defined in the registered Protocol Profile	CI bounds in S3P attestation	AAL-4
MEA-2.5	Publish epoch nonce with notary signature after epoch close to enable auditor reconstruction of all sampling decisions	Published nonce matching commitment	AAL-4
MEA-2.6	Emit S3P attestation with closed schema as defined in the registered Protocol Profile, including at minimum: epoch, violation_type, n_total, n_sampled, sampling_rate, n_violations, observed_rate, confidence_level, CI_lower, CI_upper,	Notary-signed S3P attestation	AAL-4

ID	Control	Attestation Artifact	Level
	sampling_threshold, epoch_nonce_commitment, status, and signature		

MEA-3: Third-Party Testing

Requirement: AI systems SHALL undergo independent third-party testing at regular intervals across all risk taxonomy categories.

ID	Control	Attestation Artifact	Level
MEA-3.1	Conduct third-party adversarial robustness testing at intervals defined in the operator's risk management policy, not to exceed 12 months or as defined by applicable regulation	Third-party evaluation report	AAL-2
MEA-3.2	Conduct third-party safety testing (harmful outputs, out-of-scope, hallucination, bias) at intervals defined in the operator's risk management policy, not to exceed 12 months or as defined by applicable regulation	Third-party evaluation report	AAL-2
MEA-3.3	Conduct third-party tool-call security testing at intervals defined in the operator's risk management policy, not to exceed 12 months or as defined by applicable regulation (agentic systems only)	Third-party evaluation report	AAL-2
MEA-3.4	Publish testing scope, methodology, and results summary to transparency log (redacting sensitive details)	Transparency log entry	AAL-4

NIST AI RMF: MEASURE 1.3, 2.1–2.11 | **AIUC-1:** B001, C010–C012, D002, D004

MEA-4: Pre-Deployment Testing

Requirement: AI systems SHALL undergo internal testing prior to deployment and prior to any material change.

ID	Control	Attestation Artifact	Level
MEA-4.1	Conduct pre-deployment testing covering: adversarial robustness, safety (all risk taxonomy categories), hallucination, and tool-call authorization (for agentic systems)	Test results with pass/fail criteria	AAL-2
MEA-4.2	Define material change threshold (e.g., +/-10% on evaluation metrics) requiring re-testing and re-approval	Change threshold definition in policy	AAL-2

ID	Control	Attestation Artifact	Level
MEA-4.3	Document test results and approval sign-offs before deployment proceeds	Approval records	AAL-2

NIST AI RMF: MEASURE 2.3, 2.5 | **AIUC-1:** C002, E004

10. Domain 6: RESPOND – Adaptive Control and Incident Response

Scope: Bounded, cryptographically gated response to detected violations. Maps to NIST AI RMF MANAGE and ISO 42001 Clauses 8–10.

RES-1: Cryptographically Gated Control Loop

Requirement: When the attestation system detects violations exceeding policy thresholds, adaptive control actions SHALL be cryptographically gated to prevent unauthorized or unbounded modifications.

ID	Control	Attestation Artifact	Level
RES-1.1	Aggregate verified receipts and NETATTs per epoch to compute violation metrics	Epoch metrics bundle	AAL-4
RES-1.2	Upon threshold exceedance, emit signed ControlAction specifying parameter changes (sampling_prob, queue_max, rate_limit)	ControlAction attestation	AAL-4
RES-1.3	Validate ControlAction through five cryptographic gates before application: (1) signature verification, (2) epoch currency, (3) parameter bounds, (4) co-epoch receipt for metrics, (5) co-epoch NETATT	Five-gate validation receipt	AAL-4
RES-1.4	Enforce parameter bounds: p_min <= sampling_prob <= p_max; 0 <= queue_max <= q_max; 0 <= rate_limit <= r_max. Reject ControlActions exceeding bounds regardless of signature validity	Bounded parameter attestation	AAL-4

RES-2: Incident Response

Requirement: The organization SHALL maintain and exercise AI incident response plans with attested verifiable record collection.

ID	Control	Attestation Artifact	Level
RES-2.1	Document AI failure plans for: security breaches, harmful outputs, hallucinations causing financial loss, and tool-call authorization failures	Incident response plans	AAL-1
RES-2.2	Assign accountable owner for each incident type with documented escalation criteria	Accountability matrix	AAL-2
RES-2.3	Upon incident detection, generate attestation pack: all attestation receipts, NETATT states, S3P signals, and ControlActions for the affected time period	Attestation pack with transparency log proofs	AAL-4
RES-2.4	Report critical incidents to designated authorities within required timeframes with cryptographic attestation artifacts	Incident report with attached receipts	AAL-4

NIST AI RMF: MANAGE 4.1, 4.3 | **ISO 42001:** 10.2, A.8.4 | **AIUC-1:** E001–E003

RES-3: Emergency Override ("Break Glass")

Requirement: Emergency overrides SHALL be cryptographically attested, not hidden.

ID	Control	Attestation Artifact	Level
RES-3.1	Implement emergency override requiring enhanced authentication meeting AAL-4 identity binding requirements + reason code	Override authentication attestation	AAL-4
RES-3.2	Generate override receipt with full attestation (action taken, reason code, identity, timestamp)	Override receipt in transparency log	AAL-4
RES-3.3	Automatically schedule compliance review within SLA defined in operator's policy	Review scheduling attestation	AAL-4
RES-3.4	Surface all override events in audit dashboards and risk signal feeds	Override frequency in risk signals	AAL-4

RES-4: Scoped Revocation and Circuit Breaking

Requirement: The system SHALL support scoped, time-bounded revocation or equivalent circuit breaking of specific binaries, policies, or agent identities within a tenant boundary. Revocation SHALL be designed as a circuit breaker — local, bounded, self-healing — not as a centralized kill switch.

Security principle: The revocation mechanism SHALL NOT create a centralized control plane capable of network-wide propagation. No single entity — including the attestation platform, any

notary node, or any operator — SHALL be able to trigger revocation that crosses tenant boundaries. Every revocation is tenant-scoped, gated on t-of-n notary agreement, time-bounded, and rate-limited.

ID	Control	Attestation Artifact	Level
RES-4.1	Implement tenant-scoped revocation: operators SHALL support publication of a signed Revocation Receipt revoking a specific binary_hash, policy_id, or agent identity within their own tenant boundary only. Revocation signals SHALL NOT propagate across tenant boundaries	Revocation Receipt with tenant-scoped t-of-n notary signature	AAL-4
RES-4.2	Gate revocation on t-of-n notary agreement: Revocation Receipts SHALL require the same t-of-n notary verification as full attestation. No single notary, operator, or platform entity can unilaterally trigger revocation	t-of-n gated revocation verification	AAL-4
RES-4.3	Time-bound all revocations: Revocation Receipts SHALL include an expiration (current epoch + configurable TTL, maximum: 24 hours). Expired revocations automatically reset — circuit breaker model. Permanent decommissioning requires explicit policy republication, not perpetual revocation	Time-bounded revocation with automatic reset	AAL-4
RES-4.4	Rate-limit revocation signals: maximum one revocation per policy_id per epoch. The notary network SHALL reject revocation attempts exceeding rate limits, preventing denial-of-service via revocation spam	Rate-limited revocation enforcement	AAL-4
RES-4.5	Test revocation mechanism at intervals defined in the operator's risk management policy, not to exceed 12 months or as defined by applicable regulation; attest test execution, propagation latency, automatic reset behavior, and scope containment (verify no cross-tenant effect)	Revocation test receipt with scope verification	AAL-4

RES-5: Failure Mode Declaration

Requirement: Operators SHALL declare and attest their system's default behavior when the attestation infrastructure itself becomes unavailable. The standard mandates the decision be explicit and attested, not a particular choice.

ID	Control	Attestation Artifact	Level
RES-5.1	Declare failure mode for attestation infrastructure unavailability: **fail-open** (AI system continues operating unattested) or **fail-closed** (AI system halts until attestation resumes). Publish declaration to transparency log	Failure mode declaration with transparency log proof	AAL-4
RES-5.2	For fail-open declarations: log all unattested operations locally; generate retroactive attestation receipts when the notary network resumes; report unattested duration as an explicit exposure window in risk signals. Retroactive receipts generated after fail-open periods SHALL be labeled `POST_HOC` and SHALL NOT be counted as contemporaneous attestation coverage for conformance, risk-signal reporting, or litigation reporting purposes. Post-hoc receipts are reconstruction artifacts, not contemporaneous attestation	Exposure window accounting with POST_HOC classification	AAL-4
RES-5.3	For fail-closed declarations: implement graceful degradation (queue requests, display maintenance notification, route to human fallback) rather than silent failure	Fail-closed behavior documentation and test records	AAL-2
RES-5.4	Require review of failure mode declaration at intervals defined in the operator's risk management policy with sign-off from designated risk officer	Failure mode review attestation	AAL-4

Note: For healthcare deployments, the fail-open vs. fail-closed decision is clinically material. A fail-closed AI system in an emergency department may cause harm through unavailability. A fail-open system may cause harm through unmonitored operation. OVERT does not prescribe the answer — it requires the decision to be documented, attested, and priced. [See Annex C: Design Rationale for healthcare deployment considerations.]

PART 3: AGENTIC AI CONTROLS

Part 3 defines the AI-specific execution controls required when systems invoke tools, coordinate with other agents, operate under delegated capability grants, route decisions through human approval paths, or exhibit behavioral drift. These sections provide AI-layer execution control, inter-agent boundary enforcement, capability mediation, privileged action authorization, transparency to relying parties, and behavioral anomaly monitoring for agentic workflows. Per Design Principle 6 (Security by Observation), the same inline enforcement position and tamper-evident recording that produce governance evidence also produce the detection, containment, and forensic reconstruction capabilities that security operations require. Where a control is satisfied at AAL-2 or AAL-3, the resulting claim is documentation- or operator-telemetry-grade evidence rather than cryptographically independent proof.

These controls apply to AI systems where autonomous agents execute tool calls, access external resources, and make decisions without step-by-step human oversight. They are mandatory for any system classified as "Automation" capability under IDE-1.2.

11. Tool-Call Governance

TOOL-1: Pre-Execution Policy Enforcement

Requirement: Every tool call by an AI agent SHALL be evaluated against policy and attested before execution. No tool call SHALL execute without a governance decision.

ID	Control	Attestation Artifact	Level
TOOL-1.1	Intercept all tool calls at the enforcement boundary before execution reaches the external resource	Per-call attestation receipt	AAL-4
TOOL-1.2	Evaluate tool calls against a capability policy specifying: permitted tools, permitted parameter ranges, permitted destinations, and required approval gates	Policy evaluation result in receipt	AAL-4
TOOL-1.3	Block tool calls that violate policy; generate denial receipt with policy reference and violation type	Denial receipt	AAL-4

ID	Control	Attestation Artifact	Level
TOOL-1.4	For permitted calls, generate provisional receipt before execution; upgrade to full attestation after notary validation	Three-phase receipt per Section 8	AAL-4

Architectural reference: Tool calls SHOULD be validated against information flow policies that consider the provenance and capabilities of all arguments, not just the tool name. Where the system tracks data provenance (source and allowed readers), policy checks SHOULD verify that argument capabilities permit the intended data flow.

TOOL-2: Function Authorization and Parameter Validation

Requirement: AI agents SHALL be restricted to approved functions with validated parameters.

ID	Control	Attestation Artifact	Level
TOOL-2.1	Maintain an explicit function allowlist: only approved tool functions may be invoked	Allowlist hash in policy attestation	AAL-4
TOOL-2.2	Validate function parameters against defined schemas before execution (type checking, range checking, format validation)	Parameter validation result in receipt	AAL-4
TOOL-2.3	Reject function calls with parameters outside defined bounds	Rejection receipt with parameter violation detail	AAL-4

AIUC-1: D003.1

TOOL-3: Tool-Call Rate Limiting and Circuit Breaking

Requirement: AI agent tool calls SHALL be subject to rate limits, velocity caps, and circuit breakers with attested enforcement.

ID	Control	Attestation Artifact	Level
TOOL-3.1	Enforce per-tool rate limits (calls per epoch, calls per minute)	Rate limit enforcement receipts	AAL-4
TOOL-3.2	Enforce per-session and per-user velocity caps for cumulative tool actions	Velocity enforcement receipts	AAL-4
TOOL-3.3	Implement circuit breakers: halt tool execution when error rates or violation rates exceed defined thresholds within an epoch	Circuit breaker activation receipt	AAL-4
TOOL-3.4	Track tool-call recursion depth per trace_id; terminate agent execution when depth ex-	Loop termination receipt with trace_id, depth, and termination reason	AAL-4

ID	Control	Attestation Artifact	Level
	ceeds a configurable threshold defined in deployment policy. Agents caught in retry loops (call Tool A -> error -> call Tool A) are a common failure mode. The Arbiter SHALL detect repeated identical tool calls within a trace and terminate after configurable repetition limit		

AIUC-1: D003.2

TOOL-4: Human Approval Gates

Requirement: Sensitive tool operations SHALL require explicit human approval with attested identity binding.

ID	Control	Attestation Artifact	Level
TOOL-4.1	Define which tool operations require human-in-the-loop approval (financial transactions, data deletion, external communications, privilege modifications)	Approval-required policy in attestation	AAL-4
TOOL-4.2	Gate execution pending human approval; attest approval with authenticated identity, timestamp, and action reference	Approval receipt with identity binding	AAL-4
TOOL-4.3	Implement timeout for pending approvals; attest timeout as denial if approval not received	Timeout receipt	AAL-4
TOOL-4.4	Enforce maximum approval velocity for human reviewers (configurable approvals-per-minute threshold). Approvals exceeding the velocity cap SHALL be attested as potentially fatigued and flagged for secondary review. This mitigates rubber-stamping under high volume	Approval velocity enforcement receipt	AAL-4

AIUC-1: D003.4

TOOL-5: Tool-Call Logging and Audit Trail

Requirement: All AI agent tool calls SHALL be logged with sufficient detail for retrospective analysis.

ID	Control	Attestation Artifact	Level
TOOL-5.1	Log every tool call: tool name, parameters, caller identity, timestamp, epoch, policy evaluation result, and execution outcome	Tool-call log entries	AAL-3
TOOL-5.2	Ensure tool-call logs are tamper-evident: write-once storage, cryptographic hashing of entries, sequence integrity enabling gap detection	Tamper-evident log with hash chain	AAL-4
TOOL-5.3	Attest tool-call logs at each epoch boundary with notary signature over log digest	Epoch log attestation	AAL-4

AIUC-1: D003.3, E015

11.5 MCP Server Trust Governance

The Model Context Protocol (MCP) enables AI agents to invoke tools hosted on local or remote servers. Because MCP servers mediate between the agent and external resources — databases, APIs, file systems, credentials — the trust posture of the MCP server is itself a first-class governance surface. An agent's tool-call attestation is only as strong as the trust chain to the server executing the call.

This subsection defines evidence requirements for three MCP deployment patterns: managed (vendor-hosted), custom (operator-hosted), and external (third-party-hosted). Implementations that do not use MCP or equivalent tool-hosting protocols MAY omit this subsection; the omission SHALL be declared in the conformance statement Exclusions field.

MCP-1: Managed MCP Server Posture Evidence

Requirement: When an agentic system invokes tools through a managed (vendor-hosted) MCP server, the conformant implementation SHALL attest the server's governance posture at each co-epoch boundary.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
MCP-1.1	Record the managed MCP server identity (vendor, server version, configuration hash) in the co-epoch binding at each attestation epoch	Server identity fields in co-epoch record	AAL-4
MCP-1.2	Attest the transport security state between the arbiter and the managed MCP server	Transport attestation in NETATT extension	AAL-4

ID	Control	Attestation Artifact	Level
	(TLS version, certificate identity, mutual authentication status) at each epoch		
MCP-1.3	Verify and attest the managed server's published governance metadata — including data-handling commitments, geographic jurisdiction, and sub-processor disclosures — at deployment time and upon detected change	Governance metadata receipt in transparency log	AAL-3
MCP-1.4	Attest per-call routing: for each tool call routed to a managed MCP server, the receipt SHALL identify the server instance that executed the call	Server instance identifier in per-call receipt	AAL-4

> ***Note.*** *MCP-1.3 is AAL-3 rather than AAL-4 because the governance metadata originates from the vendor's own disclosures. OVERT can attest that the metadata was retrieved, verified against a published schema, and hash-committed, but cannot independently verify the vendor's operational claims. Relying parties should treat MCP-1.3 evidence as vendor-asserted, hash-sealed metadata — not as independently verified operational posture.*

DASF: 68 | **NIST AI RMF:** GOVERN 6.1 | **ISO 42001:** A.6.2.3

MCP-2: Custom MCP Server Runtime Attestation

Requirement: When an agentic system invokes tools through a custom (operator-hosted) MCP server, the conformant implementation SHALL attest the server's runtime identity, network isolation, and per-call authorization posture.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
MCP-2.1	Include the custom MCP server binary identity (binary hash, configuration hash) in the co-epoch binding. Binary identity verification SHALL use the same mechanism as arbiter binary identity (ATT-2.2)	Server binary identity in co-epoch record	AAL-4
MCP-2.2	Attest that the custom MCP server operates within the same network isolation boundary as the arbiter, or attest the cross-boundary transport security state if it does not	Network topology attestation in NETATT	AAL-4

ID	Control	Attestation Artifact	Level
MCP-2.3	Enforce per-call authorization at the MCP server boundary: each tool invocation SHALL be evaluated against the deployment policy before execution, with the authorization decision attested in the per-call receipt	Authorization decision in per-call receipt	AAL-4
MCP-2.4	Detect and attest configuration changes to the custom MCP server within an epoch. Unauthorized configuration changes SHALL generate governance alerts with the same quality as topology change detection (MULTI-2.2)	Configuration change detection receipt	AAL-4

> ***Note.*** *MCP-2 applies operator-grade attestation to custom MCP servers because the operator controls the server lifecycle. This is stronger than MCP-1 (managed servers) because the operator can provide runtime identity evidence that a third-party vendor cannot. Implementations that co-locate the MCP server and arbiter in the same attested process may satisfy MCP-2.1 and MCP-2.2 implicitly through the arbiter's own co-epoch binding.*

DASF: 69 | **NIST AI RMF:** GOVERN 6.2, MANAGE 4.1 | **ISO 42001:** A.6.2.3, A.6.2.6

MCP-3: External MCP Connection Assurance

Requirement: When an agentic system connects to an external (third-party-hosted) MCP server, the conformant implementation SHALL attest the connection governance posture and enforce scope constraints on the external server's capabilities.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
MCP-3.1	Maintain an explicit external MCP server allowlist in the deployment policy. Connections to servers not on the allowlist SHALL be denied with attested denial receipts	Allowlist hash in policy attestation; denial receipt for unauthorized connections	AAL-4
MCP-3.2	Attest the external server's identity (endpoint URI, TLS certificate fingerprint, mutual authentication status) at each connection establishment and at each co-epoch boundary	External server identity in connection receipt	AAL-4
MCP-3.3	Enforce capability scoping for external MCP servers: the set of tools and parameters	Capability scope restriction in per-call receipt	AAL-4

ID	Control	Attestation Artifact	Level
	available through an external server SHALL be constrained to a declared subset of the server's advertised capabilities		
MCP-3.4	Apply output filtering (PRO-4) to all responses from external MCP servers before the response enters the agent context. The filtering decision SHALL be attested in the per-call receipt	External response filtering receipt	AAL-4
MCP-3.5	Record external MCP server connection lifecycle events (connect, disconnect, error, timeout) in the tamper-evident audit trail (TOOL-5) with the same attestation quality as tool-call events	Connection lifecycle entries in audit trail	AAL-4

> ***Note.*** *MCP-3 treats external MCP servers as untrusted by default. The allowlist (MCP-3.1) plus capability scoping (MCP-3.3) plus output filtering (MCP-3.4) create a defense-in-depth posture. Even if the external server is compromised, the attested scope constraints and filtering limit blast radius. MCP-3 does not and cannot attest the external server's internal security posture — that remains outside OVERT scope. What MCP-3 does attest is the* connection governance *applied at the operator's boundary.*

DASF: 70 | **NIST AI RMF:** GOVERN 6.1, MANAGE 2.4 | **ISO 42001:** A.6.2.3 | **OWASP Agentic:** #1, #4

12. Multi-Agent System Controls

MULTI-1: Inter-Agent Trust Boundaries

Requirement: In multi-agent systems, trust boundaries between agents SHALL be enforced and attested. Agents SHALL NOT inherit the trust level of peer agents.

ID	Control	Attestation Artifact	Level
MULTI-1.1	Enforce distinct policy evaluation for each agent in a multi-agent system; peer agent requests SHALL be evaluated against the same policy as external requests	Per-agent attestation receipts	AAL-4

ID	Control	Attestation Artifact	Level
MULTI-1.2	Attest the agent identity (binary hash, configuration) for each agent in the system independently	Per-agent co-epoch binding	AAL-4
MULTI-1.3	Monitor for inter-agent trust exploitation patterns (agents relaying requests to bypass restrictions). [See Annex C: Design Rationale for research basis on multi-agent trust exploitation vulnerabilities]	Anomaly detection attestation	AAL-4

MULTI-2: Agent Composition Attestation

Requirement: The composition and configuration of multi-agent systems SHALL be attested.

ID	Control	Attestation Artifact	Level
MULTI-2.1	Document and attest the agent topology: which agents exist, their roles, their communication paths, and their capability scopes	Agent topology attestation	AAL-4
MULTI-2.2	Detect and attest changes to agent topology within an epoch	Topology change detection	AAL-4

13. Capability-Based Access Control

Architectural reference: This section adapts capability-based access control principles for the attestation layer.

CAP-1: Data Provenance Tracking

Requirement: AI systems processing sensitive data SHALL track the provenance of values flowing through tool calls and enforce access policies based on provenance metadata.

ID	Control	Attestation Artifact	Level
CAP-1.1	Tag data values with provenance metadata indicating source (user, tool, AI-generated)	Provenance tracking in system design	AAL-3
CAP-1.2	Propagate provenance through transformations: if value C derives from values A and B, C inherits the provenance of both	Provenance propagation logic	AAL-3
CAP-1.3	Enforce policies based on provenance: e.g., data from untrusted sources SHALL NOT	Provenance-based policy enforcement receipts	AAL-4

ID	Control	Attestation Artifact	Level
	flow to sensitive tools without explicit authorization		

CAP-2: Architectural Separation

Requirement: AI systems making autonomous decisions SHALL architecturally separate trusted planning from untrusted data processing.

ID	Control	Attestation Artifact	Level
CAP-2.1	Planning components (which determine what actions to take) SHALL NOT directly process untrusted external data except through an attested mediation layer declared in deployment policy	Architectural documentation and validation; for Level 3 Agentic: machine-generated enforcement telemetry demonstrating mediation layer interposition; for Level 4 Agentic: independently verifiable evidence of mediation layer interposition as defined in the registered Protocol Profile	AAL-2; AAL-3 for Level 3 Agentic; AAL-4 for Level 4 Agentic
CAP-2.2	Data processing components handling untrusted input SHALL NOT have direct tool-calling capabilities	Capability restriction documentation; for Level 3 Agentic: machine-generated telemetry demonstrating capability restriction enforcement; for Level 4 Agentic: independently verifiable evidence of capability restriction as defined in the registered Protocol Profile	AAL-2; AAL-3 for Level 3 Agentic; AAL-4 for Level 4 Agentic
CAP-2.3	Data flowing from untrusted processing to trusted planning SHALL pass through structured schema validation that constrains the output space	Schema validation implementation	AAL-3

> ***Note.*** *At Level 1 and Level 2, CAP-2.1 and CAP-2.2 are AAL-2 documentation and process controls; conformance claims based on CAP-2 at those levels reflect documentation-grade evidence. At Level 3 Agentic, CAP-2.1 and CAP-2.2 require AAL-3 (machine-generated enforcement telemetry demonstrating that the architectural separation is actively enforced, not merely documented). At Level 4 Agentic, CAP-2.1 and CAP-2.2 require AAL-4 (independently verifiable evidence of architectural separation, as*

> *defined in the registered Protocol Profile — for example, hardware-attested process isolation, independently observed network segmentation, or equivalent mechanisms that do not rely solely on operator-controlled telemetry). This progressive elevation reflects the principle that evidence-grade claims about architectural separation require evidence-grade proof, not operator-controlled telemetry.*

14. Agent Disclosure and Transparency

DISC-1: Agent Transparency Documentation

Requirement: Organizations deploying agentic AI systems SHALL publish transparency documentation describing agent capabilities, constraints, and attestation status.

ID	Control	Attestation Artifact	Level
DISC-1.1	Publish agent capability documentation: which tools are available, what actions the agent can take, what constraints are enforced	Agent capability document	AAL-1
DISC-1.2	Publish AI Bill of Materials (CycloneDX AI BOM or SPDX 3.0) documenting model, components, and dependencies	AIBOM in machine-readable format	AAL-2
DISC-1.3	Publish attestation summary: coverage ratio, S3P safety signals, override frequency, and gap accounting — all derived from the attestation stream with no content exposure	Attestation summary in OSCAL format	AAL-4

AIUC-1: E017

15. Human-in-the-Loop Attestation

Human-in-the-loop interactions within AI workflows SHALL receive the same attestation quality as automated enforcement decisions. [See Annex C: Design Rationale for analysis of the verification gap in human-AI governance interactions.]

HITL-1: Consent Attestation

Requirement: When an AI system requires human consent before interaction (recording, data processing, autonomous actions affecting the individual), the consent event SHALL be attested at AAL-4 with identity binding, timestamp, and scope.

ID	Control	Attestation Artifact	Level
HITL-1.1	Define which AI interactions require prior human consent (recording, PHI processing, autonomous actions affecting the individual) and publish consent-required policy to transparency log	Consent-required policy in attestation configuration	AAL-4
HITL-1.2	Attest consent event with: authenticated identity of consenting party, timestamp, scope of consent (what was consented to), and method of consent (verbal, written, digital signature)	Consent receipt with identity binding	AAL-4
HITL-1.3	Gate AI interaction on consent receipt: the system SHALL NOT proceed with consent-required interactions without a valid consent attestation	Consent gate enforcement receipt (permit/deny)	AAL-4
HITL-1.4	Attest consent withdrawal with timestamp and scope; system SHALL cease consent-gated operations upon withdrawal attestation	Withdrawal receipt with enforcement confirmation	AAL-4

[See Annex C: Design Rationale for regulatory context on consent attestation requirements.]

HITL-2: Human Review Attestation

Requirement: When AI outputs are routed for human review (escalation, quality assurance, regulatory requirement), the review event, reviewer identity, and decision SHALL be attested at AAL-4.

ID	Control	Attestation Artifact	Level
HITL-2.1	Define which AI outputs require human review before delivery or action (clinical recommendations, financial decisions, content moderation, high-severity classifications) and publish review-required policy	Review-required policy in attestation configuration	AAL-4
HITL-2.2	Attest review event with: reviewer authenticated identity, timestamp, review decision (approve / reject / modify), and reference	Review receipt with identity binding	AAL-4

ID	Control	Attestation Artifact	Level
	to the AI output under review (by digest, not content)		
HITL-2.3	Gate output delivery or action on review receipt for review-required outputs: the AI output SHALL NOT be delivered or acted upon without a valid review attestation	Review gate enforcement receipt	AAL-4
HITL-2.4	Track and attest review latency: elapsed time from flagging to review completion, per epoch	Review latency in epoch metrics	AAL-4

HITL-3: Human Correction and Override Attestation

Requirement: When a human modifies, corrects, or overrides an AI output or recommendation (non-emergency), the intervention SHALL be attested at AAL-4.

ID	Control	Attestation Artifact	Level
HITL-3.1	Attest human corrections to AI outputs with: corrector authenticated identity, timestamp, correction type (edit, rejection, substitution), and reference to original AI output (by digest)	Correction receipt with identity binding	AAL-4
HITL-3.2	Attest non-emergency human overrides of AI recommendations with: identity, timestamp, reason category, and reference to the overridden recommendation (by digest)	Override receipt (non-emergency)	AAL-4
HITL-3.3	Aggregate correction and override rates per policy per epoch; surface as a risk signal	Correction rate in epoch metrics	AAL-4

Operational note: Elevated correction rates may indicate model degradation, domain shift, policy misalignment, or reviewer disagreement with system outputs. Sustained low correction rates are not independently sufficient to establish output quality and should be interpreted together with review quality, drift, and coverage signals.

HITL-4: Policy and Configuration Approval Attestation

Requirement: Human approvals of governance policy changes and system configuration changes SHALL be attested at AAL-4 with separation of duties enforcement.

ID	Control	Attestation Artifact	Level
HITL-4.1	Attest policy change approvals with: approver authenticated identity, timestamp, policy ver-	Policy approval receipt in transparency log	AAL-4

ID	Control	Attestation Artifact	Level
	sion transition (old hash -> new hash), and change justification category		
HITL-4.2	Attest system configuration change approvals with: approver identity, timestamp, configuration delta (by hash), and approval authority reference	Configuration change approval receipt	AAL-4
HITL-4.3	Enforce and attest separation of duties: the individual requesting a policy or configuration change SHALL NOT be the sole approver; attest both requesting and approving identities	Dual-identity approval receipt	AAL-4

NIST AI RMF: GOVERN 1.4, MANAGE 1.3 | **ISO 42001:** 5.3, 6.1.3, A.3.2 | **EU AI Act:** Article 14 (Human Oversight)

15.5 Session-Scoped Attestation

Many AI interactions are organized around sessions — bounded periods of engagement between humans and AI systems (patient encounters, clinical workflows, therapy sessions, advisory engagements, educational tutoring sessions). Session boundaries carry governance significance: consent may be scoped to a session, regulatory retention may be session-delimited, and aggregate session metrics are relevant to coverage and risk assessment. This section defines attestation requirements for session lifecycle events.

SESS-1: Session Open Attestation

Requirement: When a session-based AI interaction begins, a session_open receipt SHALL be generated attesting the session initiation.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
SESS-1.1	Generate a session_open receipt at session initiation containing: session_id (unique identifier), participant identities (authenticated per HITL identity binding requirements), session type (classification per operator's session taxonomy), and timestamp	Session open receipt with co-epoch binding	AAL-4
SESS-1.2	Include consent references in the session_open receipt linking to applicable	Session open receipt with consent attestation linkage	AAL-4

ID	Control	Attestation Artifact	Level
	HITL-1 consent attestations. Where consent was obtained prior to the session (pre-session consent), the session_open receipt SHALL reference the consent receipt attestation_id. Where consent is obtained during the session (in-session consent), the consent receipt SHALL reference the session_id		
SESS-1.3	Publish session type taxonomy to the transparency log as a machine-readable artifact. Session types SHALL be declared in the operator's governance policy and SHALL map to applicable consent requirements, retention policies, and regulatory classifications	Session type taxonomy in transparency log	AAL-4

SESS-2: Session Close Attestation

Requirement: When a session ends, a session_close receipt SHALL be generated attesting the session conclusion and summary disposition.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
SESS-2.1	Generate a session_close receipt at session termination containing: session_id (matching the session_open receipt), disposition (completed, abandoned, transferred, error, timeout, terminated), session duration, total action count within the session (tool calls, reviews, approvals, and other attested events), and timestamp	Session close receipt with co-epoch binding	AAL-4
SESS-2.2	The session_close receipt SHALL reference the session_open receipt by attestation_id, forming a verifiable session boundary pair	Session close receipt with session_open attestation_id reference	AAL-4
SESS-2.3	For sessions ending with disposition "transferred," the session_close receipt SHALL include the identity of the receiving entity (human or system) and a reference to any successor session_open receipt if available	Transfer disposition receipt with successor reference	AAL-4

SESS-3: Session Consent Binding

Requirement: Consent attestations (HITL-1) SHALL be linkable to session scope.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
SESS-3.1	Consent granted for a specific session type SHALL cover actions within sessions of that type. The consent scope field in HITL-1 receipts SHALL support session-type-scoped consent declarations	Consent receipt with session type scope	AAL-4
SESS-3.2	When consent is withdrawn mid-session (per HITL-1.4), the session SHALL either terminate (generating a session_close receipt with disposition "abandoned") or continue with reduced scope as defined in the operator's consent withdrawal policy. The consent withdrawal receipt SHALL reference the session_id	Consent withdrawal receipt with session_id reference	AAL-4

SESS-4: Session-Aggregate Signals

Requirement: Per-session summary data SHALL be reportable in epoch metrics.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
SESS-4.1	Report session-aggregate signals per epoch including at minimum: session count, average session duration, action density (average actions per session), and consent coverage rate (percentage of sessions with valid consent attestation at session open)	Session aggregate signals in epoch metrics	AAL-4
SESS-4.2	Session-aggregate signals SHALL be classified as operational signals (Annex D, Section D.2) and SHALL satisfy the signal properties defined in Section 4.6	Session signals in risk signal framework	AAL-4

SESS-5: Session Context Destruction Attestation

Requirement: When session context is destroyed (as required by policy, regulation, or operator data lifecycle management), the destruction event SHALL be attested.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
SESS-5.1	Generate a session context destruction receipt when session-scoped data (conversation history, intermediate results, session state) is destroyed. The receipt SHALL include: session_id, destruction timestamp, destruction reason (policy-mandated, regulatory-required, retention-expired, operator-initiated), and a cryptographic commitment to the data being destroyed (hash of the session content, not the content itself)	Session context destruction receipt	AAL-4
SESS-5.2	Session context destruction receipts SHALL be retained in the transparency log for the operator's full retention period, even after the session context itself is destroyed. The destruction receipt proves that context existed and was destroyed; its absence from the log after a destruction event is a conformance deviation	Destruction receipt in transparency log with retention	AAL-4

> ***Note.*** *Session-scoped attestation is applicable at Level 2 and above for systems with session-based interactions. Systems that process only stateless, independent requests without session boundaries are not required to implement this section. The determination of whether a system has "session-based interactions" is made by the operator based on the system's architecture and use context.*

NIST AI RMF: GOVERN 1.1, MANAGE 1.3 | **ISO 42001:** A.6.2.8 | **EU AI Act:** Article 12, Article 14 | **HIPAA:** 45 CFR §164.530(j) (record retention)

15.6 Agent State and Prompt Governance

Agentic AI systems that persist state across sessions (conversation memory, retrieval-augmented context, tool-call history) or operate under registered prompt artifacts (system prompts, instruction templates, chain-of-thought scaffolding) introduce governance surfaces not covered by session-scoped attestation alone. This subsection defines evidence requirements for the integrity, lineage, and governance of those surfaces.

STATE-1: Durable Agent State Sealing

Requirement: Agentic systems that persist state across session boundaries SHALL seal and attest durable state transitions so that the provenance and integrity of reused state are independently verifiable.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
STATE-1.1	At session close, compute a cryptographic commitment (hash) over the durable state snapshot that will be available to the next session. Publish the commitment to the transparency log with session-binding metadata (session_id, epoch, agent_class)	State commitment receipt in transparency log	AAL-4
STATE-1.2	At session open, verify that the loaded durable state matches the commitment published at the prior session close. Verification failure SHALL generate a state-integrity governance alert and SHALL prevent the session from proceeding until the operator resolves the discrepancy or explicitly overrides with attested justification	State verification receipt or state-integrity alert	AAL-4
STATE-1.3	Maintain a hash-chained lineage of state transitions: each state commitment SHALL reference the prior state commitment hash, enabling DAG reconstruction of the full state history for a given agent or agent class	State lineage chain in transparency log	AAL-4
STATE-1.4	Attest state mutation provenance: for each mutation to durable state within a session (memory write, context update, retrieval injection), record the source (user input, tool output, AI-generated, system-injected) and the policy evaluation result that authorized the mutation	State mutation provenance in per-action receipt	AAL-4
STATE-1.5	Enforce state access scoping: durable state SHALL be retrievable only by agent classes and sessions authorized by the deployment policy. Unauthorized state access attempts SHALL be denied and attested	State access authorization receipt or denial receipt	AAL-4

> ***Note.*** *STATE-1 does not prescribe the storage mechanism for durable state. It prescribes what must be attested about state transitions. Implementations may use vector stores, relational databases, key-value stores, or file systems — the attestation requirements are storage-agnostic. The hash-chained lineage (STATE-1.3) enables an auditor to reconstruct which state version was available to which session without accessing the state content itself.*

DASF: 72 | **NIST AI RMF:** MANAGE 2.2, GOVERN 1.3 | **ISO 42001:** 8.1, A.6.2.6

STATE-2: Prompt Artifact Registration and Binding

Requirement: Organizations deploying agentic AI systems SHALL register prompt artifacts in a governance-controlled registry and bind each agent execution to the specific prompt artifact version that governed it.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
STATE-2.1	Register all prompt artifacts (system prompts, instruction templates, chain-of-thought scaffolds, few-shot exemplars) in a versioned, hash-committed registry published to the transparency log	Prompt artifact registration receipt with content hash and version	AAL-4
STATE-2.2	At session initialization, bind the active prompt artifact version to the session attestation. The prompt artifact hash SHALL appear in the session_open receipt (§15.5)	Prompt artifact hash in session_open receipt	AAL-4
STATE-2.3	Detect and attest prompt artifact changes within a session. Mid-session prompt modification SHALL generate a governance alert and a new prompt-binding receipt	Prompt change detection receipt	AAL-4
STATE-2.4	Enforce prompt-to-action traceability: for each attested action (tool call, output generation, escalation), the receipt SHALL reference the prompt artifact version that was active when the action was authorized	Prompt artifact reference in per-action receipt	AAL-4
STATE-2.5	Require that prompt artifact registration and version changes be approved by a Qualified Risk Officer (per GOV-3.5) or equivalent governance authority declared in the deployment policy. Approval SHALL be attested with identity binding	Prompt change approval receipt with identity binding	AAL-4

> ***Note.*** *STATE-2 does not require that prompt content be disclosed in receipts or the transparency log — only the hash and version. This preserves the non-egress property: a verifier can confirm that a specific prompt version governed an execution without accessing the prompt text. Organizations that choose to disclose prompt content may do so; the standard does not require it.*

DASF: 73 | **NIST AI RMF:** GOVERN 1.1, MAP 2.1 | **ISO 42001:** 6.1.2, A.6.2.6 | **EU AI Act:** Article 13

15.7 Delegated Identity Chain Attestation

In federated deployments, the principal authorizing an agent action may not be the directly authenticated user. The action may be authorized through a chain of delegated identities: a user authenticates to an IdP, the IdP issues a token, the token is exchanged for a scoped credential, the credential is used by an orchestrator that delegates to a sub-agent. Each link in that chain is a trust decision. OVERT SHALL attest the full delegation chain so that relying parties can verify who authorized what, through which intermediaries, under which constraints.

IDENT-1: Federated Identity and Token Provenance

Requirement: Agentic systems operating under federated or delegated identity SHALL attest the full identity delegation chain from the originating principal to the executing agent.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
IDENT-1.1	Record the originating principal identity (user, service account, or workload identity) in the attestation receipt for every governed action. The identity SHALL include the identity provider, the authentication method, and the authentication timestamp	Originating principal identity in per-action receipt	AAL-4
IDENT-1.2	Record the delegation chain: for each token exchange, credential delegation, or authority transfer between the originating principal and the executing agent, attest the delegating entity, the receiving entity, the scope constraints applied at delegation, and the delegation timestamp	Delegation chain in per-action receipt	AAL-4
IDENT-1.3	Verify scope narrowing at each delegation step: each delegation SHALL narrow or	Scope verification at each delegation step	AAL-4

ID	Control	Attestation Artifact	Level
	preserve (never widen) the capability scope of the prior step. Scope widening SHALL generate a governance alert and a denial receipt		
IDENT-1.4	Attest token lifetime and revocation status: for each token or credential in the delegation chain, record the issued-at time, expiration time, and (where available) revocation-check result at the time of action authorization	Token lifecycle metadata in per-action receipt	AAL-4
IDENT-1.5	For multi-agent delegation (agent A delegates to agent B), bind the delegating agent's attestation ID (parent_attestation_id per DRIFT-1.5) to the delegation chain, enabling unified identity-and-execution DAG reconstruction	Agent delegation linkage in per-action receipt	AAL-4

> ***Note.*** *IDENT-1 does not prescribe the identity provider, token format, or federation protocol. It prescribes what must be attested about the delegation chain. Implementations using OIDC, SAML, SPIFFE, or proprietary federation protocols all satisfy IDENT-1 provided they produce the required attestation artifacts. IDENT-1.3 (scope narrowing) is the critical security property: it ensures that delegation cannot silently escalate privileges.*

DASF: 67 | **NIST AI RMF:** GOVERN 1.4, MANAGE 2.3 | **ISO 42001:** A.6.2.3 | **EU AI Act:** Article 9

16. Behavioral Drift Governance

These controls address emergent behavioral changes in agentic AI systems that occur within authorized operational bounds — situations where every individual control passes but the system's aggregate behavior drifts, cascades, or produces ungovernable complexity. Behavioral drift governance is distinct from policy violation detection (covered by PROTECT and MEASURE domains): policy violation detection identifies individual actions that breach a rule, while behavioral drift

governance detects statistically significant changes in authorized behavior patterns that may indicate systemic risk.

These controls are mandatory for any system classified as "Automation" capability under IDE-1.2 that deploys two or more interacting agents or any single agent with tool-calling capabilities.

DRIFT-1: Baseline Intent Declaration

Requirement: Agentic AI systems SHALL publish and maintain a baseline intent declaration specifying the permitted agent topology, behavioral bounds per agent class, permitted spawn relationships, model bindings, and human oversight requirements. The declaration SHALL be versioned, hash-chained, and published to the transparency log.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-1.1	Publish baseline intent declaration in machine-readable format to transparency log with cryptographic timestamp	Baseline intent declaration receipt in transparency log	AAL-4
DRIFT-1.2	Declare behavioral bounds per agent class including: permitted output distribution characteristics, permitted tool selection distributions, permitted spawn topologies, and human oversight checkpoint requirements	Behavioral bounds specification in baseline intent declaration	AAL-4
DRIFT-1.3	Version-link baseline intent declarations in the transparency log (each version references the hash of the prior version)	Hash-chained version linkage in transparency log	AAL-4
DRIFT-1.4	Require that baseline intent declaration changes be approved by a Qualified Risk Officer (per GOV-3.5) with attested separation of duties	Dual-identity approval receipt for baseline change	AAL-4
DRIFT-1.5	Publish parent-child attestation linkage requirements: every agent action receipt SHALL reference the spawning agent's attestation ID (parent_attestation_id), enabling DAG reconstruction	Parent-child attestation linkage in per-call receipts	AAL-4

NIST AI RMF: GOVERN 1.1, MAP 2.1 | **ISO 42001:** 6.1.2, A.6.2.6 | **EU AI Act:** Article 9 | **DASF:** 5.2, 9.13

DRIFT-2: Behavioral Drift Detection

Requirement: Conformant agentic systems SHALL employ sequential statistical methods to detect behavioral drift per agent class, using evaluation instruments that produce temporally stable, version-consistent measurement features. Drift detection SHALL operate on dimensions specified in the baseline intent declaration.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-2.1	Implement sequential statistical analysis (method specified in registered Protocol Profile) for detecting distribution shifts in agent behavior per agent class	Drift detection configuration in baseline intent declaration	AAL-4
DRIFT-2.2	Evaluation instruments used for drift measurement SHALL demonstrate score stability across instrument versions and cross-deployment comparability. Version stability requirements are specified in the registered Protocol Profile	Evaluation instrument version attestation	AAL-4
DRIFT-2.3	Drift detection SHALL operate per-dimension (output risk, tool selection, semantic characteristics) with independent statistical tracking per dimension	Per-dimension drift statistics in epoch metrics	AAL-4
DRIFT-2.4	Attest drift detection signals with the same co-epoch binding as enforcement receipts. Drift signals SHALL include: agent class, dimension, statistical test result, confidence level, and epoch	Drift signal receipt with co-epoch binding	AAL-4
DRIFT-2.5	Implement graduated response to drift signals: log, alert, escalate, block. Each escalation level SHALL be independently attested. The escalation ladder and thresholds SHALL be declared in the baseline intent declaration	Graduated response receipt per escalation level	AAL-4
DRIFT-2.6	Support adaptive sampling escalation triggered by drift signals — sampling rate SHALL increase when drift statistics approach declared thresholds. Escalation triggers and bounds SHALL be declared in the baseline intent declaration and attested when activated	Sampling escalation receipt with trigger evidence	AAL-4

> ***Note:*** *The standard requires drift detection capability and specifies what must be measured and attested. The specific statistical method (CUSUM, EWMA, or other sequential analysis), feature extraction architecture, and evaluation instrument design are specified in the registered Protocol Profile.*

NIST AI RMF: MEASURE 2.1–2.13, MANAGE 3.1 | **ISO 42001:** 9.1, A.6.2.6 | **EU AI Act:** Article 9, 15 | **DASF:** 5.2, 10.1

DRIFT-3: Graph Topology Governance

Requirement: Conformant multi-agent systems SHALL compute and attest graph complexity metrics for each agentic execution. When graph complexity exceeds thresholds declared in the baseline intent declaration, the system SHALL generate governance alerts with attested evidence.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-3.1	Compute graph complexity metrics per execution including at minimum: total agent count, edge count, maximum depth, and branching factor	Graph complexity metrics in execution receipt	AAL-4
DRIFT-3.2	Evaluate graph complexity against thresholds declared in the baseline intent declaration	Threshold evaluation result in execution receipt	AAL-4
DRIFT-3.3	Generate attested governance alerts when graph complexity exceeds declared thresholds, including: execution DAG summary, complexity metrics, baseline threshold, and epoch binding	Graph complexity governance alert receipt	AAL-4
DRIFT-3.4	Attest spawn authorization decisions in sub-epoch time. The mechanism for real-time spawn authorization is specified in the registered Protocol Profile. Unauthorized spawn attempts SHALL generate denial receipts with the same attestation quality as tool-call denials (TOOL-1.3)	Spawn authorization receipt or spawn denial receipt	AAL-4

> ***Note:*** *DRIFT-3.4 requires real-time spawn authorization but does not prescribe the enforcement mechanism. Protocol Profile implementations may use probabilistic data structures, allowlist lookups, or other mechanisms capable of meeting the latency requirement.*

NIST AI RMF: MANAGE 2.4 | **ISO 42001:** 8.1 | **OWASP Agentic:** #2, #3 | **DASF:** 9.13

DRIFT-4: Causal Drift Attribution

Requirement: In multi-agent systems, when behavioral drift is detected in a downstream agent, conformant Level 4 Agentic systems SHALL evaluate upstream agents for correlated drift using parent-child attestation linkages. Attribution findings SHALL be attested.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-4.1	When drift is detected in a downstream agent (per DRIFT-2), evaluate upstream agents in the attestation DAG for correlated statistical changes in the same or adjacent epochs	Upstream correlation analysis in attribution receipt	AAL-4
DRIFT-4.2	Attest attribution findings including: downstream agent class, upstream agent class, correlation evidence, attestation DAG path, and epoch range	Causal attribution receipt	AAL-4
DRIFT-4.3	When causal attribution identifies an upstream root cause, propagate the graduated response (DRIFT-2.5) to the upstream agent class	Propagated graduated response receipt	AAL-4
DRIFT-4.4	Conformant implementations SHALL employ a multi-factor attribution methodology that considers, at minimum: propagated upstream drift, local downstream drift, exogenous environmental change, combined causes, and indeterminate attribution. The attribution methodology SHALL produce attribution confidence scores quantifying the strength of evidence for each attribution factor. The specific attribution formula (e.g., PathScore) is specified in the registered Protocol Profile	Attribution methodology receipt with per-factor confidence scores	AAL-4
DRIFT-4.5	Attribution results SHALL be classified using the following taxonomy: **PROPAGATED_UPSTREAM** (drift caused by upstream agent change), **LOCAL_DOWNSTREAM** (drift caused by local agent change), **EXOGENOUS** (drift	Attribution classification receipt with taxonomy code and supporting evidence	AAL-4

ID	Control	Attestation Artifact	Level
	caused by external environmental change), **COMBINED** (multiple contributing factors identified), **INDETERMINATE** (insufficient evidence for classification). The classification SHALL be included in the attribution receipt. Where the classification is COMBINED, the receipt SHALL enumerate the contributing factors and their respective confidence scores. Where the classification is INDETERMINATE, the receipt SHALL state the reason (insufficient data, conflicting evidence, or ambiguous correlation)		

> ***Note:*** *DRIFT-4 is required for Level 4 Agentic conformance because downstream drift without upstream attribution materially limits containment and post-incident reconstruction in multi-agent systems. Simpler deployments that do not claim Level 4 Agentic conformance may still omit DRIFT-4.*

NIST AI RMF: MEASURE 2.1, MANAGE 4.1 | **ISO 42001:** 10.2

DRIFT-5: Human Oversight Quality Assessment

Requirement: Conformant systems SHALL track and attest human review quality indicators including review duration, modification rate, and consistency between review decisions and risk signals. Sustained degradation in review quality indicators SHALL trigger governance escalation.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-5.1	Track per-reviewer and per-agent-class review quality indicators: review duration (time from presentation to decision), modification rate (proportion of reviews resulting in edits, rejections, or substitutions), and risk-signal consistency (agreement between review decisions and system risk classifications)	Review quality indicators in epoch metrics	AAL-4

ID	Control	Attestation Artifact	Level
DRIFT-5.2	Attest review quality indicators per epoch with the same co-epoch binding as other governance signals	Review quality attestation receipt with co-epoch binding	AAL-4
DRIFT-5.3	Define review quality degradation thresholds in the baseline intent declaration. When review quality indicators degrade below declared thresholds (e.g., review duration dropping, modification rate declining while risk signals remain elevated), generate attested governance alerts	Review quality degradation alert receipt	AAL-4
DRIFT-5.4	Review quality indicators SHALL be reported as risk signals (see Annex D and the registered Protocol Profile for signal specifications)	Review quality in risk signal payload	AAL-4

> ***Note:*** *DRIFT-5 strengthens existing HITL-2 (Human Review Attestation) and TOOL-4.4 (approval velocity enforcement) by adding substantive quality assessment beyond mechanical timing checks. It directly supports EU AI Act Article 14's requirement that humans "properly understand the relevant capacities and limitations" of the system they oversee.*

NIST AI RMF: GOVERN 2.1 | **ISO 42001:** A.3.2 | **EU AI Act:** Article 14 | **DASF:** 8.3

16.1 Evaluator Compatibility Framework

Behavioral drift detection (DRIFT-2) depends on evaluation instruments that produce structured measurement features — governance feature vectors — which are compared across time to detect distributional shifts. When evaluator versions change (new models, updated rubrics, different scoring dimensions), the resulting feature vectors may not be comparable to those produced by the prior version. Silent reuse of detector state (baselines, thresholds, statistical accumulators) across incompatible evaluator versions produces spurious drift signals or, worse, masks genuine drift. This section defines the framework for evaluator compatibility, versioning, and state management.

EVAL-1: Governance Evaluators and Structured Verdicts

Requirement: Governance evaluators — components that produce structured verdicts and governance feature vectors within the operator trust boundary — SHALL produce outputs conforming to a closed schema with declared dimensions.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
EVAL-1.1	Evaluator outputs SHALL conform to a closed schema (no undeclared fields) with a fixed, declared set of dimensions. Each dimension SHALL have a defined name, data type, value range, and semantic description	Evaluator schema artifact in transparency log	AAL-4
EVAL-1.2	Each evaluator version SHALL publish a semantic-ordering manifest: a machine-readable declaration specifying the meaning, ordering, and interpretation of each dimension in the governance feature vector. The manifest SHALL be versioned, hash-chained, and published to the transparency log	Semantic-ordering manifest with transparency log inclusion proof	AAL-4
EVAL-1.3	Evaluator outputs SHALL include the evaluator version identifier and semantic-ordering manifest hash in every structured verdict, enabling downstream consumers to verify which evaluator produced which verdict	Evaluator version and manifest hash in verdict payload	AAL-4

EVAL-2: Compatibility Domains and Detector-State Partitioning

Requirement: Evaluator versions SHALL be organized into compatibility domains within which feature vectors are longitudinally comparable. When an evaluator version change breaks compatibility, detector state SHALL be partitioned.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
EVAL-2.1	Declare compatibility domains: a compatibility domain is a set of evaluator versions whose feature vectors are longitudinally comparable (same schema, same dimensions, same semantic ordering, compatible value ranges). The active compatibility domain SHALL be published to the transparency log	Compatibility domain declaration in transparency log	AAL-4
EVAL-2.2	When a new evaluator version breaks compatibility (different schema, different dimensions, different semantic ordering, or	Detector state partition receipt with old and new domain identifiers	AAL-4

ID	Control	Attestation Artifact	Level
	materially different calibration), detector state SHALL be partitioned: the system SHALL establish a new compatibility domain with a fresh baseline, fresh statistical accumulators, and fresh drift thresholds. Silent reuse of detector state across incompatible evaluator versions is non-conformant		
EVAL-2.3	Cross-domain drift comparison SHALL NOT be performed. Drift signals from one compatibility domain SHALL NOT be compared to or aggregated with drift signals from a different compatibility domain. Each domain maintains independent statistical history	Domain isolation attestation in drift signal receipts	AAL-4

EVAL-3: Compatibility Assessment Workflow

Requirement: Before a candidate evaluator version is activated, the system SHALL execute a compatibility assessment comparing the candidate to the active evaluator.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
EVAL-3.1	The compatibility assessment SHALL evaluate, at minimum: (a) schema conformance — the candidate produces the same set of fields as the active evaluator; (b) dimensionality — the candidate produces the same number of dimensions with the same names; (c) semantic-ordering-manifest equality — the candidate's manifest matches the active evaluator's manifest; (d) missingness behavior — the candidate handles missing or null inputs identically to the active evaluator; (e) continuity metrics — the candidate's score distributions on a held-out calibration set are within declared continuity bounds of the active evaluator's distributions; (f) calibration stability — the candidate's score-to-outcome calibration on the held-out set is within declared bounds	Compatibility assessment receipt with per-criterion results	AAL-4

ID	Control	Attestation Artifact	Level
EVAL-3.2	If the compatibility assessment determines that the candidate is compatible, the candidate MAY be activated within the existing compatibility domain. If the assessment determines incompatibility on any criterion, the candidate SHALL be activated in a new compatibility domain (per EVAL-2.2)	Compatibility determination receipt (compatible / incompatible) with criterion-level detail	AAL-4
EVAL-3.3	The compatibility assessment results SHALL be published to the transparency log before the candidate evaluator is activated in production. Activation without a published compatibility assessment is non-conformant	Pre-activation compatibility assessment in transparency log	AAL-4

EVAL-4: Evaluator Version Attestation

Requirement: The active evaluator version identifier and artifact hash SHALL be attested per-epoch.

Attestation Assurance Level: AAL-4

ID	Control	Attestation Artifact	Level
EVAL-4.1	The active evaluator version identifier, artifact hash (cryptographic digest of the evaluator binary or model artifact), and compatibility domain identifier SHALL be included in the epoch summary attestation	Evaluator version binding in epoch attestation	AAL-4
EVAL-4.2	Evaluator version changes within an epoch SHALL trigger an immediate compatibility assessment (EVAL-3) and SHALL be attested as a configuration change event (per Section 18.6)	Mid-epoch evaluator change receipt	AAL-4

> ***Note.*** *The Evaluator Compatibility Framework extends DRIFT-2.2 (evaluation instrument version stability) into a complete lifecycle framework. DRIFT-2.2 requires that evaluation instruments demonstrate score stability across versions; this section specifies how to verify, attest, and manage that stability through structured compatibility domains, assessments, and state partitioning.*

All evaluator compatibility controls in this section are normative at AAL-4 for Level 3 and Level 4 Agentic conformance claims. Systems not claiming Agentic scope are not required to implement this section.

NIST AI RMF: MEASURE 2.1–2.13 | **ISO 42001:** 9.1, A.6.2.6 | **EU AI Act:** Article 9, 15

PART 4: ATTESTATION ARCHITECTURE REQUIREMENTS

Part 4 defines the evidence trust plane required for credible AI security and governance claims. It specifies the minimum architectural properties needed for trustworthy detection, investigation, audit, and defensible response: non-egress attestation, temporal binding to runtime state, statistically reproducible measurement, third-party auditability, and preservation of records in a form suitable for later verification. These sections do not establish that a deployment is secure. They establish the conditions under which claimed control execution and observed events can be checked.

17. Non-Egress Attestation Architecture

Requirement: The attestation protocol SHALL NOT require transmission of protected content outside the operator environment. Conformant receipt-service interfaces SHALL accept only cryptographic commitments and profile-defined metadata.

17.1 All AI request/response payloads SHALL be canonicalized using deterministic encoding as specified in the registered Protocol Profile. Deterministic encoding means that two conformant encoders encoding the same logical data produce identical byte sequences. The canonicalization method SHALL be version-pinned and identified by a cryptographic hash of the encoder.

17.1.1 Numeric values in attestation envelopes SHALL be represented in a lossless format as specified by the registered Protocol Profile. The encoding SHALL ensure that numeric values are not subject to platform-dependent rounding or representation variance.

17.2 Request commitments crossing the operator's trust boundary SHALL be computed using a keyed cryptographic function with tenant-scoped keys held exclusively in the operator's key management system. Raw content digests SHALL NOT egress.

> ***Informative Note:*** *The keyed commitment construction prevents rainbow table reversal of low-entropy content (e.g., PII, SSNs) by any party with ledger access.*

17.3 Attestation artifacts (prompts, responses, policy evaluations) SHALL be stored in content-addressable storage within the operator's environment, indexed by attestation commitment, and subject to the operator's data retention policies (see Section 21).

17.4 Attestation egress SHALL be constrained to a single receipt service endpoint over mutually authenticated TLS with certificate pinning. The receipt service schema SHALL be closed (reject unknown fields) and SHALL accept only cryptographic commitments — never content.

17.5 The non-egress architecture SHOULD be designed to prevent the transmission of Protected Health Information (PHI) or other regulated content, supporting minimized compliance footprints. The applicability of data processing agreements or Business Associate Agreements remains a question of applicable law and regulatory interpretation.

17.6 Implementations SHALL conform to the non-egress specifications in the registered Protocol Profile, including envelope schemas, commitment derivations, and receipt service constraints. The requirement to comply with a registered Protocol Profile is normative. The description of Protocol Profile 1.0 in Annex B is informative — it serves as a reference implementation specification. Protocol Profile 1.0 becomes normatively binding on implementations that register compliance with it. Conformance claims SHALL identify a dated, versioned profile specification.

18. Temporal Binding and Configuration Integrity

Requirement: Every attestation receipt SHALL be cryptographically bound to the system's binary identity, network isolation state, and runtime configuration during a bounded time interval, enabling retroactive proof of what configuration was running at any attested moment.

18.1 The attestation system SHALL establish bounded time intervals (epochs) during which system configuration is attested as stable. Epoch duration SHALL be configurable (recommended: 300 seconds). Specific epoch constraints are defined in the registered Protocol Profile.

18.2 System binary identity SHALL be derived by the notary through a measurement pipeline that is (a) not controlled by the attester, (b) rooted in a hardware or cryptographic trust anchor, and (c) reproducible by an independent auditor given the measurement policy. Client-supplied identity claims are insufficient for AAL-4 conformance. The notary SHALL maintain an authoritative mapping from epoch to binary identity.

> ***Note:*** *Acceptable measurement pipelines include, but are not limited to: AWS Nitro Enclave attestation documents, Intel SGX/TDX DCAP quotes, AMD SEV-SNP attestation reports, TPM 2.0 PCR-based attestation, and hypervisor-attested or orchestrator-attested measurements. Software-based measurements from hypervisors, orchestrators, or container runtimes satisfy these properties where hardware TEE attestation is unavailable, provided the measurement source is outside the attester's administrative control. Reproducible builds and binary transparency logs provide supplementary software provenance evidence but do not by themselves satisfy runtime measurement requirements. The registered Protocol Profile specifies which mechanism(s) a conformant implementation uses.*

Notary signature constructions (whether threshold signatures, multi-signatures, or other schemes achieving the t-of-n trust property) SHALL be as defined in the registered Protocol Profile. After January 1, 2031, conformant implementations SHALL use hybrid classical + post-quantum constructions as specified in the registered Protocol Profile. Pure classical signature schemes become non-conformant after that date.

> ***Informative note:*** *The t-of-n trust property — that no single notary can forge or suppress attestations — can be realized through threshold signature schemes (a single aggregated signature, e.g., BLS) or multi-signature schemes (independent per-notary signatures verified against a t-of-n policy). Threshold schemes produce compact proofs but require distributed key generation and have limited post-quantum options. Multi-signature schemes use standard per-node signing algorithms, offer straightforward post-quantum migration via FIPS 204 (ML-DSA) or FIPS 205 (SLH-DSA), and provide transparency about which notaries participated. The Protocol Profile specifies the construction; the standard requires only the trust property.*

18.3 Network isolation state SHALL be attested at each epoch. The network isolation state hash SHALL cover, at minimum: (a) the effective egress policy, (b) the identity of the enforcement component, and (c) the TLS certificate pin set. Operators MAY include additional deployment-specific inputs such as network enforcement rules, runtime environment variables affecting AI behavior, and policy controller state. The minimum input set is specified in the registered Protocol Profile.

> ***Note — Scope of network isolation attestation:*** *The inputs attested under 18.3 are* declarative policy state *— the network policies, egress rules, and enforcement configuration that govern what the AI system is permitted to reach. Ephemeral infrastructure state (e.g., pod IP assignments, container instance identifiers, rotating service credentials) is not within scope unless the operator's measurement policy (18.2) explicitly includes it. The operator defines which specific inputs compose the network isolation*

> *state hash; the standard requires that the chosen inputs are sufficient to detect policy-level changes across epoch boundaries.*

18.4 Every attestation receipt SHALL be bound to: (a) the current epoch, (b) the notary-derived binary identity, and (c) the network isolation state hash. Receipts lacking any of these bindings SHALL be rejected.

18.5 Stale-epoch submissions (referencing a past epoch) SHALL be rejected with a deterministic error code. Bounded clock skew tolerance as defined in the registered Protocol Profile (recommended: <=2 seconds) is permitted.

18.6 Configuration drift — changes to binary identity, network state, or runtime configuration — SHALL be cryptographically detectable by comparing attestation bindings across epoch boundaries.

18.7 Implementations SHALL conform to the co-epoch binding and network attestation specifications in the registered Protocol Profile.

18.8 Within-Epoch Measurement Requirements

Cross-epoch binary identity verification (18.1–18.7) detects drift between measurement points but does not preclude a just-in-time substitution attack: an adversary may present a compliant binary at epoch-boundary measurement, execute a non-compliant binary during the epoch, and restore the compliant binary before the next measurement. To close this gap, conformant implementations SHALL ensure binary identity continuity within epochs.

18.8.1 Conformant implementations SHALL satisfy at least one of the following within-epoch measurement strategies:

(a) **Continuous remeasurement.** The implementation performs binary identity verification at a regular cadence within each epoch. The minimum within-epoch measurement frequency SHALL be specified in the registered Protocol Profile.

(b) **Per-receipt liveness proofs.** Each attestation receipt (or receipt batch, where batching is profile-defined) includes a fresh binary identity measurement binding the receipt to the binary state at the time of receipt generation. The liveness proof SHALL include a timestamp and a nonce or monotonic counter to prevent replay.

(c) **Event-driven remeasurement.** The implementation triggers an immediate binary identity verification upon any detected configuration change event, process restart, library reload, container image change, or equivalent mutation to the execution environment.

Implementations that rely solely on epoch-boundary measurement without any within-epoch strategy are NOT conformant with AAL-3 or AAL-4.

18.8.2 At AAL-4, binary identity verification SHALL occur at minimum once per receipt batch (where batch size is defined by the Protocol Profile) or upon any detected configuration change event, whichever is more frequent. The measurement result SHALL be cryptographically bound to the receipt or receipt batch it covers.

18.8.3 Any gap in within-epoch measurement exceeding the profile-defined maximum interval SHALL cause the affected receipts to be marked with the status `MEASUREMENT_GAP` and SHALL be disclosed in the epoch summary.

18.8.4 Implementations MAY use hardware-rooted continuous attestation mechanisms (e.g., runtime TCB measurement via TPM, Intel TXT, or ARM TrustZone) to satisfy within-epoch measurement requirements with higher assurance. Hardware-rooted continuous attestation meeting or exceeding the Protocol Profile minimum frequency SHALL be considered sufficient without additional software-layer remeasurement.

18.8.5 The registered Protocol Profile SHALL declare the within-epoch measurement strategy, minimum measurement frequency, maximum batch size for per-receipt proofs, and the set of configuration change events that trigger event-driven remeasurement.

19. Statistical Safety Measurement

Requirement: Safety monitoring SHALL produce quantified statistical statements with exact confidence intervals, derived from cryptographically unbiased sampling that is auditor-reproducible without content access.

19.1 Sampling for AI system monitoring SHALL be deterministic and auditor-reproducible. The operator SHALL NOT be able to selectively monitor favorable interactions — an auditor SHALL be able to verify that sampling was fair and comprehensive.

19.2 Sampling decisions SHALL be cryptographically unpredictable during the observation period and verifiable after it. A secret value (nonce) SHALL drive sampling decisions during each epoch, then be published after epoch close for independent reconstruction. This standard defines a single normative auditor-reproducible sampling and measurement method: the Statistical Safety Signal Protocol (S3P). Alternative sampling constructions SHALL be specified in a registered Protocol Profile and demonstrated to preserve completeness verification and auditor reconstruction.

19.3 Safety claims SHALL carry exact confidence intervals computed using conservative statistical methods requiring no distributional assumptions (e.g., exact binomial intervals). Unquantified safety

assertions are not attestation artifacts. [See Annex B: Protocol Profile Reference Summary for formula specifications.]

19.4 Per-epoch safety attestations SHALL include at minimum: total requests, sampled count, violation count, sampling rate, observed violation rate, confidence level, confidence interval bounds, and sampling methodology identifier.

19.5 An auditor SHALL be able to verify safety claims by: (a) obtaining published epoch secrets, (b) recomputing sampling decisions for all requests, (c) verifying sample-set membership, and (d) recomputing confidence intervals — all without accessing protected content.

19.6 Implementations SHALL conform to the statistical safety signal specifications in the registered Protocol Profile.

19.7 Signal Volume Prerequisites

19.7.1 Clopper–Pearson exact binomial confidence intervals require a minimum number of sampled events before the resulting upper bound on violation rate constitutes a credible statistical claim. Implementations SHALL NOT report a violation-rate bound tighter than the sample size supports. The following table specifies minimum sample sizes for common confidence and bound combinations with zero observed violations:

Target Upper Bound	Confidence Level	Min. Sampled Events (zero violations)
10% (0.1)	95%	29
5% (0.05)	95%	59
1% (0.01)	95%	299
0.5% (0.005)	95%	598
0.1% (0.001)	95%	2,995
1% (0.01)	99%	459
0.1% (0.001)	99%	4,603

> ***Note:*** *These values assume zero observed violations. Any observed violation invalidates a zero-violation bound; the Clopper–Pearson interval then widens per standard exact binomial computation.*

19.7.2 S3P attestations generated from epochs or aggregation windows where the number of sampled events falls below the minimum required for the claimed bound SHALL report the status code `ERR_INSUFFICIENT_SAMPLE` as defined in the registered Protocol Profile. Signal consumers SHALL NOT extrapolate a violation-rate bound from an epoch or aggregation window carrying `ERR_INSUFFICIENT_SAMPLE` status.

19.7.3 Deployments where the expected sampled-event volume within a single epoch is insufficient to meet the minimum sample size for the target bound SHALL use longer aggregation windows. Conformant approaches include:

(a) **Extended aggregation windows.** The implementation MAY aggregate sampled events over longer periods (e.g., daily, weekly). The aggregation period SHALL be explicitly disclosed in every S3P attestation produced under this mode.

(b) **Rolling windows.** The implementation MAY use a rolling window of the most recent n_min sampled events, provided the window boundary timestamps are included in the attestation.

Implementations SHALL NOT silently default to epoch-level bounds when volume is insufficient.

19.7.4 The coverage ratio reported in S3P attestations SHALL identify its denominator source and SHALL reference independently verifiable ingress metrics where available (e.g., load balancer request counts, API gateway telemetry), rather than relying solely on DPL completeness metrics measured inside the declared mediation scope. Where independent ingress metrics are not available, the attestation SHALL disclose this limitation and SHALL mark the denominator as operator-declared. Level 4 claims SHALL use independently verifiable ingress metrics or a registered-Protocol-Profile equivalent denominator source.

20. Third-Party Auditability

Requirement: The attestation system SHALL enable third-party verification of AI governance claims without requiring trust in the operator or access to protected content.

20.1 All attestation receipts SHALL be recorded in an append-only transparency log conformant with RFC 6962 (Certificate Transparency) that provides: (a) inclusion proofs (receipt exists in log), (b) consistency proofs (log was not modified between time points), and (c) split-view detection (operator cannot show different logs to different auditors).

20.2 Machine-readable attestation packs SHALL be expressible in standard compliance formats (e.g., OSCAL Assessment Results) for interoperability with existing audit workflows.

20.3 Auditor verification SHALL be possible at three levels: (a) sampling integrity verification (using published ledgers and epoch secrets), (b) configuration integrity verification (using co-epoch bindings), and (c) content verification (accessing verifiable records in operator's local storage under legal authority).

20.4 Routine verification (levels a and b) SHALL operate entirely on cryptographic artifacts without content access. Content verification (level c) SHALL be the exception, used only under legal authority or contractual agreement, with cryptographic proof that accessed attestation artifacts are genuine and contemporaneous.

20.5 Implementations SHALL conform to the auditor verification procedures in the registered Protocol Profile.

21. Legal Preservation and Production

Requirement: The attestation architecture SHALL define retention, preservation, export, and chain-of-custody requirements sufficient to support regulatory examination and litigation discovery, without compromising non-egress guarantees or cryptographic verifiability.

21.1 Retention Requirements

Operators SHALL define and publish a retention schedule for each attestation artifact class (receipts, attestation packs, S3P signals, ControlActions), mapped to applicable legal, regulatory, and contractual requirements. The retention schedule SHALL be attested in the transparency log. Operators are responsible for determining the retention periods appropriate to their jurisdictions and regulatory environment.

21.2 Legal Hold

Upon receipt of a litigation hold notice, preservation demand, or regulatory investigation notice, the operator SHALL:

(a) Suspend automated deletion of all attestation artifacts within scope.

(b) Generate a point-in-time attestation pack with transparency log inclusion proofs for all in-scope receipts.

(c) Attest the legal hold activation with timestamp and scope definition.

21.3 Immutable Export

Operators SHALL be capable of producing an immutable export package containing:

(a) All attestation receipts for a defined time period and scope.

(b) Transparency log inclusion and consistency proofs.

(c) Notary signatures and epoch data.

(d) S3P attestations and ControlActions.

(e) Custodian certification (identity of export operator, timestamp, scope declaration, hash of export package).

The export package SHALL be independently verifiable as to integrity, signature validity, and transparency-log consistency by any party with access to the transparency log and published epoch data. Production of operator-local artifacts (e.g., full CAS records) may require operator cooperation or lawful process.

21.4 Chain of Custody

Each attestation artifact SHALL include sufficient metadata to establish chain of custody: creation timestamp, creator identity (notary or arbiter), epoch binding, and transparency log position.

21.5 Retroactive Receipt Classification

The receipt `flags` field SHALL be a fixed-width unsigned integer. Bit 0 (least significant) indicates attestation temporality:

- **0** = contemporaneous: the attestation was generated and attested within the same epoch as the governed event.
- **1** = POST_HOC: the attestation was generated during the governed event but attested in a subsequent epoch (for example, due to network unavailability during a fail-open period per RES-5.2).

Remaining bits are reserved and SHALL be set to zero. The field width is specified in the registered Protocol Profile.

POST_HOC receipts are reconstruction artifacts and SHALL NOT be counted as contemporaneous attestation coverage for conformance, risk-signal reporting (Annex D), or litigation reporting purposes. POST_HOC receipts SHALL be distinguishable from contemporaneous receipts in all export packages, signal computations, and audit reports.

> ***Note — Scope of POST_HOC classification:*** *POST_HOC applies only when notary co-signing is delayed across an epoch boundary — that is, the governed event occurred in epoch* N *but the notary signature was obtained in epoch* N+1 *or later. Transient network latency* within *an epoch does not trigger POST_HOC classification. Under the three-phase attestation model (ATT-3), Phase 1 (policy*

> *enforcement) and Phase 2 (provisional receipt generation) execute locally and synchronously; only Phase 3 (notary co-signature upgrade) is asynchronous. Because the recommended epoch duration (18.1) is 300 seconds, routine network micro-outages are absorbed without reclassification.*

21.6 Redaction Procedures

When attestation packs must be produced with certain content redacted (e.g., to protect third-party PHI in multi-tenant environments), redaction SHALL preserve the cryptographic verifiability of unredacted portions. Redacted fields SHALL be replaced with their cryptographic commitments, enabling a verifier to confirm that the redacted content was present without accessing it. Redacted fields SHALL be salted with an operator-held secret prior to commitment generation. The salt SHALL be unique per field per attestation to prevent dictionary inversion of low-entropy content.

[See Annex C: Design Rationale for legal admissibility considerations and the role of attestation architecture in litigation readiness.]

PART 5: CONFORMANCE AND CROSSWALKS

22. Conformance

22.1 Overview

OVERT conformance is expressed as a composite claim combining a **maturity level** (1–4) and a **scope designator** (Core, Agentic, or Agentic-Extended). That claim describes the depth of control-execution evidence an implementation produces within its declared scope; it does not constitute a general representation that the system is secure, compliant, or safe.

OVERT conformance requires AAL-4 attestation for all controls designated as AAL-4 in this standard. Controls designated AAL-1, AAL-2, or AAL-3 require the specified level. Conformance is assessed per-control, not globally.

The maturity level determines which governance domains, attestation architecture requirements, and response or preservation capabilities are in scope for verification. The scope designator determines whether the claim is limited to non-agentic operation or extends to agentic execution paths such as tool use, inter-agent coordination, delegated capability use, disclosure, drift governance, MCP trust governance, durable state governance, prompt registration, and delegated identity attestation.

Every conformance claim at any level SHALL include a human-readable scope summary identifying the systems, interfaces, and traffic classes covered, and a human-readable exclusions summary identifying what is not covered. Level 1 and Level 2 claims SHALL include the scope summary and exclusions summary. Level 3 and Level 4 claims SHALL additionally identify the mediation scope statement hash, the declared coverage percentage of the mediation scope relative to its denominator, the denominator source used for coverage and measurement claims, whether that denominator source is independently verifiable or operator-declared only, and the total exposure-window duration (periods of unattested operation) during the claim period. A Level 3 or Level 4 claim that includes optimistic enforcement SHALL also disclose the percentage of in-scope actions processed under

optimistic enforcement during the claimed period, including both the claim-period average and the worst single-epoch value.

A conformant implementation SHALL state its conformance using the grammar defined in Section 22.4 and SHALL satisfy every normative requirement associated with its claimed level and scope.

22.2 Maturity Levels

OVERT defines four cumulative maturity levels. Each level subsumes all requirements of the preceding level.

Level	Name	Governance Domains (Part 2)	Attestation Architecture (Part 4)	Summary
1	Foundation	GOVERN (Section 5): GOV-1 – GOV-5; IDENTIFY (Section 6): IDE-1, IDE-2	None	Documented governance basis and system characterization. Buyers and auditors may verify that policies, inventories, and impact assessments were documented. Runtime enforcement, continuous monitoring, and incident-grade evidence are outside Level 1.
2	Enforcement	Level 1 + PROTECT (Section 7): PRO-1 – PRO-5; HITL (Section 15): HITL-1, HITL-4	Section 17 (Non-Egress Architecture), Section 18 (Temporal Binding)	Adds attested boundary enforcement, non-egress architecture, temporal binding, and defined human approval paths. Buyers and auditors may verify that declared execution controls operated at the runtime boundary for in-scope actions.
3	Measurement	Level 2 + ATTEST (Section 8): ATT-1 – ATT-4; MEASURE (Section 9): MEA-1 – MEA-4; HITL (Section 15): HITL-2, HITL-3	Level 2 + Section 19 (Statistical Measurement), Section 20 (Auditability)	Adds independently useful telemetry, statistical measurement, transparency-log auditability, and full human-review evidence. Auditors and defenders may verify sampling integrity, coverage disclosures, measurement

Level	Name	Governance Domains (Part 2)	Attestation Architecture (Part 4)	Summary
				outputs, and attested review events for in-scope operations.
4	Evidence-Grade	Level 3 + RESPOND (Section 10): RES-1 – RES-5; ATTEST (Section 8): ATT-5	Level 3 + Section 21 (Legal Preservation)	Adds attested response actions, IAP governance, and preservation or export controls for later investigation. This is the highest OVERT evidence and preservation tier. It does not establish overall security, regulatory compliance, insurer endorsement, or entitlement to insurance coverage.

> ***Note.*** *HITL controls (Section 15) appear in Part 3 but are required at Core scope because human oversight obligations arise for both agentic and non-agentic systems at Levels 2–4. HITL is architecturally situated in Part 3 for editorial coherence with the agentic control family, not because it is agentic-only.*

22.3 Scope Designators

The scope designator controls whether an implementation must additionally satisfy the agentic-specific control families in Part 3 (Sections 11–16, with HITL excluded from the scope gate). HITL (Section 15) is required by the maturity level regardless of scope.

Scope	Sections Required	Description
Core	Part 2 (Sections 5–10) per level + Section 15 (HITL) per level + Part 4 (Sections 17–21) per level	Systems that do not autonomously invoke external tools, coordinate with other agents, or operate under delegated authority.
Agentic	Core + TOOL (Section 11, excluding Section 11.5) + MULTI (Section 12) + CAP (Section 13) + DISC (Section 14) + DRIFT (Section 16), each per level	Systems that autonomously invoke external tools, participate in multi-agent orchestration, operate under delegated capability grants, or exhibit potential for goal drift. Does not use MCP or equivalent tool-

Scope	Sections Required	Description
		hosting protocols for external tool invocation.
Agentic-Extended	Agentic + MCP (Section 11.5) + STATE (Section 15.6) + IDENT (Section 15.7), each per level	Agentic systems that additionally invoke tools through MCP servers (managed, custom, or external), persist durable agent state across session boundaries, register prompt artifacts, or operate under federated/delegated identity chains.

The controls required at each level are:

Level	Agentic Controls (in addition to Core)	Agentic-Extended Controls (in addition to Agentic)
1	None (policy and inventory only)	None (policy and inventory only)
2	TOOL-1 – TOOL-5, DRIFT-1, DRIFT-3.4	MCP-1 (if managed MCP), MCP-3.1 (if external MCP), STATE-2.1
3	Level 2 + MULTI-1 – MULTI-2, CAP-1 – CAP-2, DISC-1, DRIFT-2, DRIFT-3, DRIFT-5	Level 2 Extended + MCP-1 – MCP-3, STATE-1, STATE-2, IDENT-1.1 – IDENT-1.3
4	Level 3 + DRIFT-4	Level 3 Extended + IDENT-1.4, IDENT-1.5

An implementation deploying agentic capabilities SHALL claim Agentic scope. Claiming Core scope for a system that autonomously invokes external tools or coordinates with other agents is non-conformant regardless of maturity level.

An implementation deploying agentic capabilities that use MCP servers, persist durable agent state, register prompt artifacts, or operate under federated identity SHALL claim Agentic-Extended scope. Claiming Agentic scope (without the Extended qualifier) for a system that uses MCP servers or persists durable agent state is non-conformant. Where only a subset of the Agentic-Extended control families applies, the conformance statement Exclusions field SHALL declare the omitted family and the architectural justification.

> ***Note.*** *At Level 1, the Agentic scope designator indicates only that the system deploys agentic capabilities and that the operator has satisfied the Level 1 documentation requirements. It does not indicate that agentic-specific enforcement, monitoring, or drift controls are in place.*

> ***Note.*** *At Level 3 Agentic, CAP-2.1 and CAP-2.2 are elevated to AAL-3 (machine-generated enforcement telemetry). Claims about architectural separation based on CAP-2 at Level 3 therefore reflect opera-*

> *tor-controlled telemetry-grade evidence, not cryptographically independent proof.* ***Level 3 Agentic conformance statements SHALL explicitly state that CAP-2 evidence is AAL-3 (operator-controlled telemetry) in the conformance claim itself, not only in supporting documentation.*** *At Level 4 Agentic, CAP-2.1 and CAP-2.2 require AAL-4 (independently verifiable evidence as defined in the registered Protocol Profile). Level 4 Agentic claims about architectural separation therefore require evidence beyond operator-controlled telemetry. Conformance claims at Level 4 Agentic that cannot satisfy AAL-4 for CAP-2 SHALL NOT assert evidence-grade architectural separation.*

22.4 Conformance Statement Grammar

A conformance claim SHALL take one of the following forms:

```
For Level 1 and Level 2:
OVERT Level <N> <Scope> – <Standard-Version>, <Profile-Version>, Scope Summary:
<Scope-Summary>, Exclusions: <Exclusions-Summary>, [ABD: <ABD-Hash>,] <Date>

For Level 3 and Level 4:
OVERT Level <N> <Scope> – <Standard-Version>, <Profile-Version>, Scope Summary:
<Scope-Summary>, Exclusions: <Exclusions-Summary>, Scope: <Coverage-Percent> of
<Denominator-Description>, Denominator: <Independent|Operator-Declared>, Scope
Statement: <Scope-Hash>, Exposure Window: <Exposure-Duration>, [Optimistic:
<Optimistic-Average>/<Optimistic-Worst-Epoch> of in-scope actions,] IAP Topology:
<Single-IAP|Multi-IAP>, [Arbiter Isolation: Software-Only,] [ABD: <ABD-Hash>,]
<Date>
```

Where:

- `<N>` is the maturity level (1, 2, 3, or 4).
- `<Scope>` is `Core`, `Agentic`, or `Agentic-Extended`.
- `<Standard-Version>` is the OVERT standard version (e.g., `v1.0.0`).
- `<Profile-Version>` is the registered protocol profile version (e.g., `Profile v1.0`). For Level 1, where no protocol profile is operationally required, this field SHALL read `No Profile` or reference the intended target profile.
- `<Scope-Summary>` is a human-readable summary enumerating the system identifiers, interfaces, and traffic classes covered by the claim. The scope summary SHALL enumerate specific system identifiers (not generic descriptions), the interfaces through which attested traffic flows, and the traffic classes within scope. Generic or free-text-only scope summaries are non-conformant.
- `<Exclusions-Summary>` SHALL take one of three forms: (1) `None (full coverage verified)` — all identified in-scope traffic and interfaces are

attested; (2) `Not assessed: <list>` — identified systems or interfaces that have not yet been evaluated for conformance, enumerated by identifier; (3) a specific exclusion list with per-item justification stating why each excluded item is outside the claim scope. Free-text exclusion summaries without per-item justification are non-conformant for Level 3 and Level 4 claims.

- `<Coverage-Percent>` is the declared mediation-scope coverage percentage for the claim.
- `<Denominator-Description>` is a human-readable description of the denominator used for the coverage claim (e.g., `inbound API traffic`).
- `<Independent|Operator-Declared>` states whether the denominator source is independently verifiable or operator-declared only.
- `<Scope-Hash>` is the mediation scope statement hash identifying the published scope artifact.
- `<Exposure-Duration>` is the total duration of unattested operation (exposure windows) during the claim period, expressed as hours and as a percentage of the claim period. If zero, this field SHALL read `0h (0%)`.
- `<Optimistic-Average>` is the claim-period average percentage of in-scope actions processed under optimistic enforcement.
- `<Optimistic-Worst-Epoch>` is the worst single-epoch optimistic enforcement percentage during the claim period.
- `IAP Topology: <Single-IAP|Multi-IAP>` is mandatory for ALL Level 4 claims. Both Single-IAP and Multi-IAP deployments SHALL declare their IAP topology.
- `Arbiter Isolation: Software-Only` is included when Section 4.7.3(f) requires disclosure that the AAL-4 arbiter is not running in a hardware-attested TEE.
- `ABD: <ABD-Hash>` is mandatory for Agentic-Extended claims and identifies the published Attestation Boundary Declaration defined in Section 29.4.
- `<Date>` is the ISO 8601 date on which the conformance assessment was completed.

Examples:

```
OVERT Level 2 Core – v1.0.0, Profile v1.0, Scope Summary: sys-cda-001 clinical
documentation API (FHIR R4 interface, HL7v2 ADT feed), Exclusions: Not assessed:
batch-analytics-002 (scheduled for Q3 assessment), 2026-03-15
OVERT Level 3 Agentic – v1.0.0, Profile v1.0, Scope Summary: sys-agent-010 patient-
facing agentic workflows (API gateway gw-prod-01, FHIR interface, voice endpoint),
Exclusions: None (full coverage verified), Scope: 85% of inbound API traffic,
Denominator: Independent, Scope Statement: sha256:<scope-hash>, Exposure Window: 0h
(0%), IAP Topology: Multi-IAP, 2026-02-28
OVERT Level 4 Agentic-Extended – v1.0.0, Profile v1.0, Scope Summary: sys-agent-010
and sys-cds-020 production agentic workflows (API gateway gw-prod-01, internal RPC
mesh, FHIR R4 interface), Exclusions: None (full coverage verified), Scope: 100% of
declared in-scope actions, Denominator: Independent, Scope Statement: sha256:<scope-
hash>, Exposure Window: 2h (0.03%), Optimistic: 8%/22% of in-scope actions, IAP
```

```
Topology: Single-IAP, ABD: sha256:<abd-hash>, 2026-03-15
OVERT Level 1 Core — v1.0.0, No Profile, Scope Summary: sys-ambient-005 ambient
clinical documentation system (voice capture endpoint, EHR integration interface),
Exclusions: Not assessed: sys-transcribe-006 non-AI transcription workflows,
2026-01-10
```

> ***Note.*** *Conformance claims are point-in-time assertions. A conformance claim does not represent ongoing conformance unless accompanied by continuous attestation evidence at Level 3 or above. Implementations SHOULD include the standard version and profile version in all conformance documentation.*

22.5 Conformance Matrix

The following matrix maps the primary control-family requirements for each Level–Scope combination. This matrix is non-exhaustive: conformance additionally requires satisfaction of the normative overlays in Sections 4.1 (AAL mapping), 4.5 (threat model mitigations), 4.6 (risk signal properties and verifiability classification), 4.7 (security considerations including IAP compromise response, log monitor diversity, arbiter hardening, mediation scope attestability, and anomaly triage), 4.8 (cross-boundary attestation for cross-boundary workflows), 22.1 (scope and exclusions disclosure), 22.6 (protocol profile registration), 22.7 (IAP qualification), and 22.8 (qualified assessor requirements). All applicable normative overlays SHALL be satisfied for the claimed level. A cell marked **R** indicates the requirement is mandatory (SHALL). A cell marked **S** indicates the requirement is recommended (SHOULD). A cell marked **—** indicates the requirement does not apply. All requirements are cumulative: Level N includes all requirements from Levels 1 through N−1.

Section	Control Family	L1 Core	L1 Agentic	L2 Core	L2 Agentic	L3 Core	L3 Agentic	L4 Core	L4 Agentic
Part 2									
§5	GOVERN (GOV-1 – GOV-5)	R	R	R	R	R	R	R	R
§6	IDENTIFY (IDE-1, IDE-2)	R	R	R	R	R	R	R	R
§7	PROTECT	—	—	R	R	R	R	R	R

Section	Control Family	L1 Core	L1 Agentic	L2 Core	L2 Agentic	L3 Core	L3 Agentic	L4 Core	L4 Agentic
	(PRO-1 – PRO-5)								
§8	ATTEST (ATT-1 – ATT-4)	—	—	—	—	R	R	R	R
§8	ATTEST (ATT-5)	—	—	—	—	—	—	R	R
§9	MEA-SURE (MEA-1 – MEA-4)	—	—	—	—	R	R	R	R
§10	RE-SPOND (RES-1 – RES-5)	—	—	—	—	—	—	R	R
Part 3									
§11	TOOL (TOOL-1 – TOOL-5)	—	—	—	R	—	R	—	R
§12	MULTI (MULTI-1 – MULTI-2)	—	—	—	—	—	R	—	R
§13	CAP (CAP-1 – CAP-2)	—	—	—	—	—	R	—	R
§14	DISC (DISC-1)	—	—	—	—	—	R	—	R
§15	HITL (HITL-1, HITL-4)	—	—	R	R	R	R	R	R

Section	Control Family	L1 Core	L1 Agentic	L2 Core	L2 Agentic	L3 Core	L3 Agentic	L4 Core	L4 Agentic
§15	HITL (HITL-2, HITL-3)	—	—	—	—	R	R	R	R
§15.5	SESS (SESS-1 – SESS-5)	—	—	R	R	R	R	R	R
§16	DRIFT (DRIFT-1, DRIFT-3.4)	—	—	—	R	—	R	—	R
§16	DRIFT (DRIFT-2, DRIFT-3, DRIFT-5)	—	—	—	—	—	R	—	R
§16	DRIFT (DRIFT-4)	—	—	—	—	—	—	—	R
§16.1	EVAL (EVAL-1 – EVAL-4)	—	—	—	—	—	R	—	R
Part 4									
§17	Non-Egress Architecture	—	—	R	R	R	R	R	R
§18	Temporal Binding	—	—	R	R	R	R	R	R
§19	Statistical Measurement	—	—	—	—	R	R	R	R
§20	Auditability	—	—	—	—	R	R	R	R
§21	Legal Preservation	—	—	—	—	—	—	R	R

Section	Control Family	L1 Core	L1 Agentic	L2 Core	L2 Agentic	L3 Core	L3 Agentic	L4 Core	L4 Agentic
Part 1									
§4.8	Cross-Boundary Attestation	—	—	—	—	R	R	R	R

> ***Note.*** *§15.5 (SESS) applies at Level 2+ for systems with session-based interactions; systems without session-based interactions are exempt. §16.1 (EVAL) applies at Level 3+ Agentic. §4.8 (Cross-Boundary Attestation) applies at Level 3+ for cross-boundary workflows; single-boundary deployments are exempt. §22.8 (Qualified Assessor) applies as: Level 3 SHOULD, Level 4 SHALL.*

Agentic-Extended Overlay

Agentic-Extended claims add the following control-family overlays on top of the base Agentic matrix:

Control	Level 1	Level 2	Level 3 Agentic-Extended	Level 4 Agentic-Extended
MCP-1.1	—	Required (if managed MCP)	Required	Required
MCP-1.2	—	Required (if managed MCP)	Required	Required
MCP-1.3	—	Required (if managed MCP)	Required	Required
MCP-1.4	—	Required (if managed MCP)	Required	Required
MCP-2.1	—	—	Required (if custom MCP)	Required
MCP-2.2	—	—	Required (if custom MCP)	Required
MCP-2.3	—	—	Required (if custom MCP)	Required
MCP-2.4	—	—	Required (if custom MCP)	Required
MCP-3.1	—	Required (if external MCP)	Required	Required
MCP-3.2	—	—	Required (if external MCP)	Required
MCP-3.3	—	—	Required (if external MCP)	Required
MCP-3.4	—	—	Required (if external MCP)	Required
MCP-3.5	—	—	Required (if external MCP)	Required
STATE-1.1	—	—	Required	Required

Control	Level 1	Level 2	Level 3 Agentic-Extended	Level 4 Agentic-Extended
STATE-1.2	—	—	Required	Required
STATE-1.3	—	—	Required	Required
STATE-1.4	—	—	Required	Required
STATE-1.5	—	—	Required	Required
STATE-2.1	—	Required	Required	Required
STATE-2.2	—	—	Required	Required
STATE-2.3	—	—	Required	Required
STATE-2.4	—	—	Required	Required
STATE-2.5	—	—	Required	Required
IDENT-1.1	—	—	Required	Required
IDENT-1.2	—	—	Required	Required
IDENT-1.3	—	—	Required	Required
IDENT-1.4	—	—	—	Required
IDENT-1.5	—	—	—	Required

22.6 Protocol Profile Registry Governance

A protocol profile defines the specific cryptographic primitives, serialization formats, and transport bindings that satisfy the normative profile-dependent clauses throughout the standard. Profile-dependent clauses appear in Part 2 (Sections 5–10) for schema definitions, governance artifact formats, and measurement output structures, and in Part 4 (Sections 17–21) for cryptographic algorithms, envelope structures, hash functions, signature schemes, key hierarchies, and evidence serialization.

An implementation claiming OVERT Level 2 or above SHALL reference a registered protocol profile. The profile SHALL cover all profile-dependent normative clauses applicable to the claimed level — not solely those in Part 4.

22.6.1 Registry Publication

The Protocol Profile Registry SHALL be published at a stable URL with complete version history. Each registry entry SHALL include the profile identifier, version, submission date, registration date, and a persistent link to the full profile specification.

22.6.2 Submission and Registration

Any party MAY submit a protocol profile for registration. The registry maintainer SHALL accept or reject submissions within 90 calendar days of receipt. A submission SHALL satisfy:

1. **Normative coverage.** The profile SHALL specify concrete cryptographic constructions, envelope schemas, key derivation methods, and receipt formats satisfying every normative SHALL requirement applicable to the claimed scope.
2. **Test vectors.** The profile SHALL include published test vectors for every cryptographic operation.
3. **Deterministic verification.** Given identical inputs and the profile specification, any two independent implementations SHALL produce identical cryptographic outputs.
4. **Public specification.** The profile specification SHALL be publicly available for inspection.
5. **Patent disclosure.** The profile submission SHALL disclose any known patent claims that may be essential to implementation.
6. **Conformance test suite.** The submission SHALL include or reference a publicly available conformance test suite sufficient to verify implementation correctness.

Evaluation is mechanical: profiles meeting all criteria SHALL be registered. The registry maintainer SHALL NOT reject profiles on grounds other than failure to meet the criteria above.

22.6.3 Self-Declaration Upon Registry Non-Response

If the registry maintainer fails to issue a written acknowledgment within 14 calendar days, or fails to render a decision within 90 calendar days of receipt of a complete submission, the submitter MAY publish the profile as a self-declared profile. A self-declared profile SHALL include a prominent notice stating that registry registration was attempted but no response was received, and SHALL use the profile identifier prefix `SD-` to distinguish it from registry-registered profiles. A self-declared profile is valid for Level 1 and Level 2 conformance claims. A self-declared profile SHALL NOT be used for Level 3 or Level 4 conformance claims. Level 3 and Level 4 claims require a registry-registered profile because the evidence-grade and measurement-grade claims at those levels depend on third-party-reviewed cryptographic constructions, test vectors, and conformance test suites that self-declaration cannot provide.

22.6.4 Registry Governance Policy

The registry governance policy SHALL be published alongside the registry and SHALL specify criteria for acceptance and rejection, appeals process, update and deprecation procedures, registry maintainer identity and contact information, and succession conditions. Changes to the governance policy SHALL be published with at least 60 calendar days advance notice.

22.6.5 Registry Continuity

If the registry maintainer ceases operations for more than 180 consecutive calendar days, any organization MAY establish a successor registry provided it incorporates all entries from the prior

registry, publishes a governance policy meeting the requirements of Section 22.6.4, and provides at least 90 calendar days public notice before accepting new submissions.

22.7 Independent Attestation Provider (IAP) Qualification

An entity operating as an Independent Attestation Provider (IAP) per Section 3.14 SHALL satisfy the following requirements:

Structural independence:

1. The IAP SHALL NOT hold equity in, be a subsidiary of, or share common management with the AI system operator whose attestations it validates.
2. The IAP SHALL maintain contractual independence: the operator SHALL NOT have unilateral authority to suppress, modify, or delay attestation artifacts.
3. The IAP SHALL disclose any material business relationships with operators whose attestations it validates.
4. The IAP SHALL disclose its beneficial ownership structure to operators upon request.

Operational requirements:

5. The IAP SHALL publish uptime and availability metrics for its notary infrastructure.
6. The IAP SHALL provide auditor access to epoch nonces, digest publication ledgers, and transparency log entries as required by Section 20.
7. The IAP SHALL maintain key management practices consistent with the security requirements of the registered Protocol Profile.
8. The IAP SHALL demonstrate operational capability for all Part 4 sections required by the highest level it services.

Transparency and accountability:

9. The IAP SHALL publish its operational policies, including key management practices, geographic distribution of notary infrastructure, and incident response procedures.
10. The IAP SHALL disclose any security incidents affecting attestation integrity within 72 hours of detection (see Section 4.7.1).
11. The IAP SHALL publish a transparency report at least annually, covering receipt volume, verification failure rates, and any compromise or coercion events to the extent permitted by law.

Portability and resilience:

12. Operators SHALL be able to transition between IAPs without loss of historical attestation data. The outgoing IAP SHALL provide transparency log entries and published epoch data for the transition period.

13. The IAP SHALL support portability escrow: upon operator request, the IAP SHALL export all configuration artifacts, epoch data, and transparency log entries necessary for a replacement IAP to assume attestation services. The export format SHALL be documented and publicly specified.
14. For Level 4 claims, the IAP SHALL cooperate with the operator's annual migration rehearsal (Section 4.7.1(g)) by providing a test environment or equivalent mechanism sufficient to validate the portability escrow.

Any entity meeting these requirements MAY operate as an IAP. This standard does not restrict IAP operation to any specific commercial entity.

22.8 Qualified OVERT Assessor Program

This section defines the requirements for third-party assessors who evaluate OVERT conformance on behalf of operators, relying parties, or regulatory bodies.

22.8.1 Purpose

A Qualified OVERT Assessor is a third-party entity that independently evaluates an operator's OVERT conformance claim against the normative requirements of this standard. The Qualified Assessor program establishes minimum competence, independence, and procedural requirements for assessment activities, ensuring that conformance claims at higher maturity levels are subject to rigorous independent evaluation.

22.8.2 Assessor Requirements

An entity seeking qualification as an OVERT Assessor SHALL satisfy the following requirements:

Independence:

(a) The assessor SHALL be structurally independent of the assessed organization. The same structural independence requirements applicable to IAPs (Section 22.7, items 1–4) apply to assessors, adapted for assessment: the assessor SHALL NOT hold equity in, be a subsidiary of, or share common management with the organization whose conformance it assesses. The assessor SHALL maintain contractual independence and disclose any material business relationships with assessed organizations.

(b) The assessor SHALL not have provided implementation consulting, system design, or protocol profile development services to the assessed organization for the system under assessment within the 24 months preceding the assessment. This does not preclude prior training or educational engagements.

Competence:

(c) The assessor SHALL demonstrate competence in: (i) OVERT standard interpretation — documented understanding of all normative requirements across Parts 1–5, including version-specific changes; (ii) cryptographic verification procedures — ability to independently verify receipt signatures, co-epoch bindings, transparency log inclusion and consistency proofs, and S3P attestation recomputation; (iii) attestation infrastructure assessment — ability to evaluate deployment topology, arbiter isolation, notary governance, and mediation scope completeness; (iv) governance framework evaluation — documented understanding of the governance frameworks crosswalked in Sections 23–29 sufficient to evaluate OVERT conformance claims in context.

(d) The assessor SHALL maintain a documented assessment methodology aligned with this standard, including checklists, evidence collection procedures, and report templates.

Professional standards:

(e) The assessor SHALL maintain professional liability coverage adequate to the scope of assessments performed.

(f) The assessor's assessment staff SHALL complete annual continuing education in AI governance and attestation, with documented training records. A minimum of 16 hours of continuing education per year is RECOMMENDED.

22.8.3 Assessment Procedures

Qualified Assessors SHALL conduct assessments using the following procedures:

(a) **Scope validation.** Verify that the claimed conformance scope (systems, interfaces, traffic classes) matches the actual deployment. The assessor SHALL independently confirm that the systems identified in the conformance statement are the systems under attestation, and that the mediation scope statement accurately describes the attested traffic.

(b) **Control verification.** Test each claimed control against normative requirements at the claimed level. For AAL-4 controls, the assessor SHALL independently verify at least one representative attestation artifact per control family (receipt signature, co-epoch binding, transparency log proof). For AAL-1 through AAL-3 controls, the assessor SHALL review the documented evidence.

(c) **Evidence review.** Verify attestation artifacts against transparency log entries. The assessor SHALL independently retrieve receipts from the transparency log, verify inclusion proofs, and confirm that the artifacts presented by the operator match the log entries.

(d) **Signal validation.** Independently recompute risk signals from published data for at least one representative epoch. The assessor SHALL verify that the operator's reported coverage ratio, violation rate bounds, and gap accounting are consistent with the published epoch data and transparency log entries.

(e) **Report generation.** Produce a standardized assessment report per Section 22.8.4.

22.8.4 Assessment Reports

Assessment reports SHALL include:

(a) **Assessed organization and system identification.** Legal entity name, system identifiers, deployment environment description, and attestation infrastructure description (IAP identity, protocol profile, deployment topology).

(b) **Claimed conformance level and scope.** The operator's conformance statement (per Section 22.4 grammar) as claimed.

(c) **Assessment date range and methodology version.** The start and end dates of the assessment period, the OVERT standard version assessed against, and the assessor's methodology version.

(d) **Per-control findings.** For each control applicable to the claimed level and scope: conformant, non-conformant, or not applicable. Non-conformant findings SHALL include a description of the deficiency and the normative requirement not satisfied.

(e) **Identified deficiencies and recommended remediation.** A summary of all non-conformant findings with specific remediation recommendations and, where applicable, a recommended timeline for remediation.

(f) **Assessor certification and signature.** The lead assessor's identity, the assessor organization's identity, the assessor's qualification status (registry identifier per Section 22.8.5), and a signed certification that the assessment was conducted in accordance with this standard and the assessor's documented methodology.

22.8.5 Assessor Registry

GLACIS Technologies or successor registry maintainer SHALL maintain a public registry of Qualified OVERT Assessors. The registry SHALL include: assessor organization identity, qualification date, qualification scope (which maturity levels and scope designators the assessor is qualified to assess), and annual requalification status. Registry governance SHALL follow the same continuity provisions as the Protocol Profile Registry (Section 22.6.5).

22.8.6 Level Requirements for Assessment

(a) Level 1 and Level 2 conformance MAY be self-assessed by the operator.

(b) Level 3 conformance SHOULD use a Qualified OVERT Assessor. Self-assessment at Level 3 SHALL be disclosed in the conformance statement.

(c) Level 4 conformance SHALL use a Qualified OVERT Assessor. Level 4 conformance claims not supported by a Qualified Assessor's assessment report are non-conformant.

(d) A managed deployment, hosted reference implementation, or implementation-vendor attestation service MAY improve deployment assurance and evidence readiness, but SHALL NOT be represented as equivalent to Qualified Assessor certification unless the assessment is performed by an entity satisfying the independence requirements of this section.

22.9 Envelope Format vs. Conformance Claims

Systems MAY produce attestation artifacts using the OVERT envelope format (defined in Section 17 and detailed in Annex B) without making a conformance claim. Use of the envelope format indicates structural compatibility with OVERT tooling and verification procedures, but does not constitute an assertion that the system satisfies the normative requirements of any Attestation Assurance Level.

An OVERT conformance claim requires satisfaction of all normative requirements at the claimed AAL level, assessment by a qualified assessor (where required by Section 22.8.6), and disclosure per Section 22.3.

23. Crosswalk: NIST AI RMF

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with the NIST AI Risk Management Framework.

This crosswalk maps OVERT controls to the NIST AI Risk Management Framework (AI RMF 1.0, January 2023) and the NIST AI RMF Generative AI Profile (July 2024).

NIST AI RMF Function / Category	OVERT Domain	OVERT Controls	Coverage	Notes
GOVERN 1.1–1.7 (Policies and procedures)	GOVERN, DRIFT	GOV-1, GOV-2, DRIFT-1	Partial	Machine-readable policy attestation and baseline intent declaration; OVERT attests policy existence and cryptographic binding but does not establish, implement, or evaluate policy adequacy

NIST AI RMF Function / Category	OVERT Domain	OVERT Controls	Coverage	Notes
GOVERN 2.1–2.3 (Accountability structures)	GOVERN, DRIFT	GOV-2, DRIFT-5	Partial	Role and responsibility attestation; human oversight quality assessment. OVERT attests accountability structures but does not establish them
GOVERN 3.1–3.2 (Workforce diversity and training)	GOVERN	GOV-2	Adjacent	OVERT attests organizational role assignments but does not address workforce diversity, composition, or training programs
GOVERN 4.1–4.3 (Organizational commitments)	GOVERN	GOV-5	Partial	Disclosure and transparency attestation supports evidence of organizational commitments
GOVERN 5.1–5.2 (Engagement)	GOVERN	GOV-4	Adjacent	OVERT attests supply chain governance records; stakeholder engagement processes, community consultation, and feedback mechanisms are outside OVERT scope
GOVERN 6.1–6.2 (Policies and procedures review)	GOVERN	GOV-1	Partial	Policy review cycle attestation provides evidence that reviews occurred; review quality and adequacy are outside scope
MAP 1.1–1.6 (Context and use identification)	IDENTIFY	IDE-1	Partial	System inventory and classification attestation; context analysis and intended-use documentation are organizational responsibilities
MAP 2.1–2.3 (Context assessment)	IDENTIFY	IDE-1.2, IDE-2	Partial	Impact categorization and risk assessment attestation; contextual factors and deployment environment analysis are organizational responsibilities
MAP 3.1–3.5 (Benefits and costs)	IDENTIFY	IDE-1, IDE-2	Adjacent	Risk documentation supports evidence; benefit-cost analysis and societal impact assessment are outside OVERT scope
MAP 4.1–4.2 (Risk identification)	IDENTIFY	IDE-2, GOV-4	Partial	Risk registry with attestation; risk identification methodology and completeness are organizational responsibilities
MAP 5.1–5.2 (Stakeholder impacts)	IDENTIFY	IDE-2	Adjacent	Impact assessment attestation provides supporting evidence; stakeholder identification and impact analysis are outside OVERT scope
MEASURE 1.1–1.3 (Metrics identification)	MEASURE	MEA-1, MEA-2	Direct	S3P sampling infrastructure produces attested metrics

NIST AI RMF Function / Category	OVERT Domain	OVERT Controls	Coverage	Notes
MEASURE 2.1–2.13 (AI system evaluation)	MEASURE, DRIFT	MEA-2, MEA-3, MEA-4, DRIFT-2	Partial	Statistical safety signals with confidence intervals; behavioral drift detection per agent class. OVERT provides runtime measurement; comprehensive AI system evaluation including fairness, bias, and societal impact assessment requires additional evaluation methods
MEASURE 3.1–3.3 (Tracking and communication)	MEASURE, ATTEST	MEA-2, ATT-4	Direct	Transparency log integration provides attested tracking and communication records
MEASURE 4.1–4.3 (Measurement feedback)	MEASURE	MEA-3	Partial	TEVV process attestation supports evidence of feedback loops; measurement methodology adequacy is outside scope
MANAGE 1.1–1.4 (Risk response planning)	RESPOND	RES-1, RES-2	Partial	Attested response actions with cryptographic receipts; risk response planning, strategy development, and resource allocation are organizational responsibilities
MANAGE 2.1–2.4 (Risk treatment)	RESPOND, PROTECT, DRIFT	RES-1, PRO-1, DRIFT-3	Partial	Enforcement attestation; graph topology governance. OVERT attests that treatments executed but does not evaluate treatment effectiveness
MANAGE 3.1–3.2 (Risk monitoring)	MEASURE, DRIFT	MEA-1, MEA-2, DRIFT-2	Partial	Continuous monitoring via S3P and behavioral drift governance within attested scope; risk monitoring scope completeness and adequacy are organizational responsibilities
MANAGE 4.1–4.3 (Risk escalation)	RESPOND, DRIFT	RES-2, RES-3, DRIFT-4	Partial	Escalation and override attestation; causal drift attribution within attested scope. Escalation path design and organizational decision-making are outside scope
GOVERN 1.4 (Policies updated)	HITL	HITL-4	Direct	Policy approval attestation
MANAGE 1.3 (Risk responses)	HITL	HITL-2, HITL-3	Direct	Human review and correction attestation

24. Crosswalk: ISO/IEC 42001:2023

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with ISO/IEC 42001:2023.

This crosswalk maps OVERT controls to ISO/IEC 42001:2023 (Artificial Intelligence — Management System) clauses and Annex A controls.

ISO 42001 Clause / Annex	OVERT Domain	OVERT Controls	Coverage	Notes
4.1–4.4 Context of the organization	IDENTIFY	IDE-1	Adjacent	OVERT system inventory supports evidence of organizational context documentation; understanding of the organization, interested parties, and AIMS scope are management responsibilities
5.1–5.3 Leadership	GOVERN	GOV-1, GOV-2	Partial	Policy commitment and organizational role attestation provide evidence of leadership engagement; leadership commitment, resource allocation, and management review are outside OVERT scope
6.1.1–6.1.3 Risk assessment	IDENTIFY, GOVERN, DRIFT	IDE-2, GOV-3, DRIFT-1	Partial	AI risk classification with attestation; baseline intent declaration. OVERT supports evidence of risk assessment execution but does not prescribe risk assessment methodology or criteria
6.1.4 AI system impact assessment	IDENTIFY	IDE-2	Partial	Severity classification attestation supports evidence that impact assessments occurred; assessment completeness and methodology are organizational responsibilities
6.2 AI objectives	GOVERN	GOV-1.1	Adjacent	Objective-level attestation provides supporting records; establishing AI objectives and planning to achieve them is a management responsibility
7.5 Documented information	ATTEST	ATT-1, ATT-4	Partial	Tamper-evident records in transparency log address integrity and protection of documented information; OVERT does not address the full documented-information lifecycle including creation, updating, document

ISO 42001 Clause / Annex	OVERT Domain	OVERT Controls	Coverage	Notes
				control scope, and retention of all management system records
8.1 Operational planning and control	PROTECT	PRO-1 through PRO-5	Partial	Boundary enforcement attestation produces verifiable evidence of operational control execution within attested scope; operational planning, resource allocation, and control selection are organizational responsibilities
8.2–8.4 AI risk assessment and treatment	IDENTIFY, RESPOND	IDE-2, RES-1	Partial	Risk treatment attestation supports evidence of treatment execution; risk assessment completeness and treatment selection are organizational responsibilities
9.1 Monitoring, measurement, analysis	MEASURE, DRIFT	MEA-1, MEA-2, DRIFT-2	Direct	S3P statistical measurement; behavioral drift detection produces the monitoring artifacts
9.2 Internal audit	ATTEST	ATT-4	Adjacent	Transparency log provides audit-ready infrastructure and tamper-evident records; OVERT does not conduct internal audits, establish audit programs, define audit criteria, or ensure auditor independence — these are management system obligations
10.2 Corrective action	RESPOND, DRIFT	RES-2, DRIFT-4	Partial	Attested corrective actions; causal drift attribution. OVERT attests that corrective actions were taken but does not evaluate their adequacy or confirm nonconformity elimination
A.6.2.4 Verification and validation	MEASURE	MEA-3, MEA-4	Direct	TEVV attestation
A.6.2.6 Operation and monitoring	MEASURE, PROTECT, DRIFT	MEA-2, PRO-1, DRIFT-1, DRIFT-2	Direct	Runtime monitoring attestation; baseline intent and drift governance
A.6.2.8 Event log recording	ATTEST, TOOL	ATT-1, TOOL-5	Direct	Per-action attestation receipts
A.3.2 Roles and responsibilities	HITL, DRIFT	HITL-4.3, DRIFT-5	Partial	Separation of duties attestation; human oversight quality assessment. OVERT attests role assignments but does not define organizational roles or ensure competence
A.7.5 Data provenance	Agentic Controls	CAP-1	Direct	Capability-based data access attestation with provenance tracking

25. Crosswalk: EU AI Act

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with Regulation (EU) 2024/1689.

This crosswalk maps OVERT controls to Regulation (EU) 2024/1689 (Artificial Intelligence Act). The Regulation generally applies from **2 August 2026**. Article 6(1) obligations and corresponding provisions apply from **2 August 2027**. Annex III systems are classified under Article 6(2) and follow the general application date. References in this crosswalk reflect the Regulation as adopted.

OVERT attestation architecture is designed to support the requirements of the EU AI Act but does not determine conformity. Conformity assessment for high-risk AI systems under Article 43 is conducted by notified bodies or through internal control procedures as applicable.

EU AI Act Article	Subject	OVERT Domain	OVERT Controls	Coverage	Notes
Article 9	Risk management system	GOVERN, IDENTIFY, DRIFT	GOV-3, IDE-2, DRIFT-1, DRIFT-2	Partial	OVERT attests risk classification execution and provides behavioral drift detection, supporting evidence that risk monitoring occurred. Article 9 requires a complete risk management system lifecycle — including risk identification, estimation, evaluation, elimination/mitigation of risks, and testing — that extends well beyond runtime attestation
Article 10	Data and data governance	PROTECT	PRO-5, CAP-1	Partial	Data boundary enforcement attestation; capability-scoped data access. Article 10 data governance requirements (training data quality, representativeness, bias examination) are outside OVERT scope
Article 11	Technical documentation	ATTEST, GOVERN	ATT-1, ATT-4, GOV-1	Partial	Tamper-evident technical documentation in transparency log; machine-readable policy records. OVERT supports integrity and availability of documentation but does not generate the full technical documentation required by Annex IV

EU AI Act Article	Subject	OVERT Domain	OVERT Controls	Coverage	Notes
Article 12	Record-keeping	ATTEST, TOOL	ATT-1 through ATT-5, TOOL-5	Partial	Automatic logging with cryptographic attestation; tamper-evident records; retroactive reconstruction via transparency log. OVERT provides attested automatic logging records aligned with Article 12 record-keeping objectives, but Article 12 may require additional system-specific data elements outside OVERT scope
Article 13	Transparency and provision of information to deployers	GOVERN	GOV-5, DISC-1	Partial	AI system disclosure attestation; deployer-facing transparency records. OVERT supports evidence of transparency measures but does not generate all deployer-facing information required by Article 13
Article 14	Human oversight	HITL, DRIFT	HITL-1 through HITL-4, DRIFT-5	Partial	Human-in-the-loop attestation: consent, review, correction, and override with cryptographic receipts; human oversight quality assessment. OVERT attests that oversight occurred but does not ensure oversight effectiveness or design appropriate oversight measures
Article 15	Accuracy, robustness, and cybersecurity	PROTECT, MEASURE, DRIFT	PRO-1 through PRO-4, MEA-2, MEA-3, DRIFT-2	Partial	Boundary enforcement attestation; statistical safety measurement with confidence intervals; behavioral drift detection. OVERT provides runtime evidence but does not address the full accuracy, robustness, and cybersecurity lifecycle including design-time measures
Article 17	Quality management system	GOVERN, MEASURE	GOV-1, GOV-2, MEA-4	Adjacent	Governance policy attestation; quality management process records. Article 17 requires a comprehensive QMS covering design, development, testing, and post-market obligations; OVERT provides supporting attestation evidence for runtime aspects only
Article 26	Obligations of deployers	ATTEST, HITL	ATT-4, HITL-1, HITL-4	Partial	Deployer logging obligations; human oversight documentation. OVERT supports evidence of deployer obligations but does not address all deployer responsibilities under Article 26

EU AI Act Article	Subject	OVERT Domain	OVERT Controls	Coverage	Notes
Article 72	Reporting of serious incidents	RESPOND	RES-1, RES-2, RES-5	Partial	Attested incident detection and response; declared failure modes. OVERT provides incident detection evidence; the reporting obligation itself and causal investigation are organizational responsibilities

26. Crosswalk: AIUC-1 / OWASP

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with AIUC-1 or OWASP guidance.

This crosswalk maps OVERT controls to the AIUC-1 framework (January 2026 version) and the OWASP Top 10 for Agentic Applications, showing how OVERT attestation infrastructure supports evidence for these requirements by upgrading assurance levels over documentation-based approaches.

AIUC-1 / OWASP Reference	OVERT Domain	OVERT Controls	Coverage	Attestation Assurance Level Upgrade
A001–A007 (Data and Privacy)	PROTECT	PRO-5	Partial	Documentation to AAL-4 attestation receipts for data boundary enforcement; data privacy program design and DPIAs are outside scope
B001 (Adversarial Testing)	MEASURE	MEA-3.1	Partial	OVERT attests that testing occurred and publishes results to transparency log; OVERT does not conduct adversarial testing
B004 (Endpoint Protection)	PROTECT	PRO-3	Direct	Documentation to AAL-4 rate limit attestation receipts
B006 (Agent Actions)	TOOL, PROTECT	TOOL-1, PRO-1	Direct	Documentation to AAL-4 per-action attestation receipts
B008 (Deployment Security)	ATTEST	ATT-2	Direct	Documentation to AAL-4 co-epoch binding
C001 (Risk Taxonomy)	GOVERN	GOV-3	Partial	Document to AAL-4 machine-readable taxonomy; OVERT attests taxonomy existence and

AIUC-1 / OWASP Reference	OVERT Domain	OVERT Controls	Coverage	Attestation Assurance Level Upgrade
				binding but does not evaluate taxonomy correctness or completeness
C003–C005 (Safety Controls)	PROTECT	PRO-4	Direct	Documentation to AAL-4 filter attestation receipts
C010–C012 (Third-party Testing)	MEASURE	MEA-3	Direct	Reports to AAL-2 + AAL-4 transparency log summary
D003 (Tool Call Restriction)	TOOL	TOOL-1 through TOOL-5	Direct	Documentation to AAL-4 per-call attestation
E015 (Logging)	ATTEST, TOOL	ATT-1 through ATT-4, TOOL-5	Direct	Logs to AAL-4 tamper-evident attested records
E016 (AI Disclosure)	GOVERN	GOV-5	Partial	Demonstrations to AAL-2 disclosure attestation; disclosure content adequacy is outside scope
E004 (Accountability)	HITL	HITL-4	Direct	Document to AAL-4 policy approval receipts
D003.4 (Approval Gates)	HITL	HITL-1, HITL-2, HITL-3	Direct	Documentation to AAL-4 consent/review/correction receipts
OWASP Agentic #1 (Prompt Injection)	PROTECT, TOOL	PRO-1, TOOL-1	Partial	Boundary enforcement attestation at tool-call boundary; prompt injection prevention effectiveness depends on filter quality, which OVERT attests but does not determine
OWASP Agentic #2 (Tool Misuse)	TOOL, DRIFT	TOOL-1 through TOOL-5, DRIFT-3	Direct	Per-call policy evaluation with attestation receipt; graph topology governance
OWASP Agentic #3 (Privilege Escalation)	Agentic Controls, DRIFT	CAP-1, CAP-2, DRIFT-3	Partial	Capability-based access control with attestation; spawn authorization governance. CAP-2 architectural separation evidence is AAL-3 at Level 3 and AAL-4 at Level 4; privilege escalation prevention effectiveness depends on policy quality and scope completeness

27. Crosswalk: NIST SP 800-53 Rev 5 / FedRAMP

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with NIST SP 800-53 Rev 5 or FedRAMP requirements.

This crosswalk maps OVERT controls to NIST SP 800-53 Revision 5 security and privacy control families. Organizations pursuing FedRAMP authorization at the Moderate or High baseline can use this mapping to identify how OVERT attestation architecture supports evidence for required control implementations.

OVERT does not replace 800-53 controls. It provides a cryptographic attestation layer that supports evidence for audit, integrity, and accountability controls by producing independently verifiable records of control execution.

800-53 Family	Representative Controls	OVERT Controls	Coverage	Alignment Notes
AU — Audit and Accountability	AU-2 (Event Logging), AU-3 (Content of Audit Records), AU-6 (Audit Record Review), AU-9 (Protection of Audit Information), AU-10 (Non-repudiation), AU-11 (Audit Record Retention), AU-12 (Audit Record Generation)	ATT-1 through ATT-5, TOOL-5, Section 21	Partial	Strong alignment for AI governance audit records within attested scope. OVERT attestation receipts produce tamper-evident, timestamped, independently verifiable audit records. Transparency log provides AU-9 (protection via append-only cryptographic structure). Section 21 addresses AU-11 (retention). ATT-2 (co-epoch binding) supports AU-10 (non-repudiation). OVERT covers AU controls for AI governance events; system-wide AU controls (non-AI event logging, centralized audit reduction, cross-domain audit) require separate implementation.
SI — System and Information Integrity	SI-4 (System Monitoring), SI-5 (Security Alerts), SI-6 (Security and Privacy Func-	PRO-1 through PRO-5, MEA-1 through MEA-4,	Partial	OVERT boundary enforcement (PRO controls) provides attested system integrity monitoring for AI governance controls. S3P (MEA-2) produces statistical safety measurements with confidence intervals. Co-epoch binding (ATT-2) supports SI-7 by attesting binary and configuration in-

800-53 Family	Representative Controls	OVERT Controls	Coverage	Alignment Notes
	tion Verification), SI-7 (Software, Firmware, and Information Integrity)	DRIFT-1 through DRIFT-3		tegrity. DRIFT controls provide behavioral drift detection (SI-4), graph topology governance alerts (SI-5), and baseline intent verification (SI-6). OVERT addresses AI-layer integrity; host-level and network-level SI controls require separate implementation.
IA — Identification and Authentication	IA-2 (Identification and Authentication), IA-3 (Device Identification and Authentication), IA-8 (Identification and Authentication — Non-Organizational Users), IA-12 (Identity Proofing)	HITL-1 through HITL-4, ATT-2	Partial	OVERT HITL controls provide attested identity binding for human actors in AI governance workflows. ATT-2 provides cryptographic identity binding for system components via notary-derived binary identity. OVERT does not implement authentication mechanisms, identity proofing, or credential management — these require separate IA control implementations.
CM — Configuration Management	CM-2 (Baseline Configuration), CM-3 (Configuration Change Control), CM-6 (Configuration Settings), CM-8 (System Component Inventory)	ATT-2 (co-epoch binding), Section 18	Partial	Co-epoch binding attests configuration state at each epoch. Configuration drift is cryptographically detectable (Section 18.6). Binary identity and network isolation state are independently verified by notaries. OVERT provides configuration integrity evidence for AI governance components; CM controls for the broader system environment require separate implementation.
IR — Incident Response	IR-4 (Incident Handling), IR-5 (Incident Monitoring), IR-6 (Incident Reporting), IR-8 (Incident Response Plan)	RES-1 through RES-5	Partial	OVERT RESPOND controls provide cryptographically gated incident response with attested escalation, override, and revocation procedures. RES-5 (declared failure modes) supports IR-8 planning requirements. OVERT addresses AI governance incidents; organization-wide IR program, staff training, and IR plan development are outside scope.

800-53 Family	Representative Controls	OVERT Controls	Coverage	Alignment Notes
AC — Access Control	AC-3 (Access Enforcement), AC-4 (Information Flow Enforcement), AC-6 (Least Privilege), AC-25 (Reference Monitor)	CAP-1, CAP-2, TOOL-2	Partial	OVERT capability-based access control (CAP controls) implements least privilege with attestation at AI tool-call boundaries. TOOL-2 (schema enforcement) acts as a reference monitor for tool-call boundaries. Information flow enforcement is attested via non-egress architecture. OVERT provides access control attestation for AI system actions; system-wide AC controls (user account management, session controls, remote access) require separate implementation.
SC — System and Communications Protection	SC-7 (Boundary Protection), SC-8 (Transmission Confidentiality and Integrity), SC-12 (Cryptographic Key Establishment and Management), SC-13 (Cryptographic Protection)	PRO-2, ATT-1, Section 17	Partial	Non-egress architecture (Section 17) provides boundary protection with attestation for AI governance data flows. ATT-1 defines cryptographic commitment constructions. Split-knowledge key hierarchy supports SC-12. OVERT addresses SC controls for attestation infrastructure; network-level boundary protection and system-wide encryption require separate implementation.
SA — System and Services Acquisition	SA-4 (Acquisition Process), SA-9 (External System Services), SA-11 (Developer Testing and Evaluation)	GOV-1, MEA-3, ATT-5	Adjacent	OVERT governance policy attestation supports evidence for SA-4 acquisition documentation. MEA-3 (TEVV) provides attested testing and evaluation. ATT-5 (notary governance) documents external service dependencies. OVERT provides supporting evidence; SA controls address broader acquisition lifecycle and vendor management processes outside OVERT scope.
PM — Program Management	PM-9 (Risk Management Strategy), PM-28 (Risk Framing)	GOV-3, IDE-2	Adjacent	OVERT risk classification (GOV-3) and impact assessment (IDE-2) produce attested risk management records that support evidence for PM-level risk strategy documentation. Risk management strategy development and organizational risk framing are management responsibilities outside OVERT scope.

FedRAMP Considerations

For cloud service providers (CSPs) pursuing FedRAMP authorization:

- OVERT attestation architecture can serve as supporting evidence for AU and SI control implementations in the System Security Plan (SSP). OVERT does not satisfy these controls independently.
- Transparency log data supports evidence for continuous monitoring requirements under the FedRAMP Continuous Monitoring Strategy Guide.
- The non-egress architecture (Section 17) is designed to minimize transfer of customer content outside the operator boundary. Classification of attestation artifacts within a specific authorization boundary is determined by the system security plan and the authorizing official.
- S3P statistical safety signals provide quantitative metrics suitable for Plan of Action and Milestones (POA&M) tracking.

28. Crosswalk: OMB M-25-21 / M-25-22

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with OMB M-25-21, M-25-22, or any other federal requirement.

This crosswalk maps OVERT controls to federal AI procurement and governance requirements established by Office of Management and Budget memoranda M-25-21 ("Accelerating Federal Use of AI through Innovation, Governance, and Public Trust," April 2025) and M-25-22 ("Driving Efficient Acquisition of Artificial Intelligence in Government," effective for solicitations issued on or after **October 1, 2025** (180 days after issuance)).

Important: M-25-22 excludes National Security Systems (NSS) from its scope. Organizations operating NSS should consult applicable national security directives and their authorizing officials for AI governance requirements.

M-25-21: Agency AI Governance and Risk Management

M-25-21 requires federal agencies to inventory AI use cases, implement risk management practices, and report on AI governance (initial reporting targeting **April 2026** per OMB instructions). OVERT controls support evidence for these requirements as follows:

M-25-21 Requirement	OVERT Domain	OVERT Controls	Coverage	Alignment Notes
AI use case inventory and registration	GOVERN, IDENTIFY	GOV-1, GOV-5, IDE-1	Partial	Machine-readable governance policy (GOV-1) and system inventory (IDE-1) produce attested records that support evidence for AI use case registration. GOV-5 disclosure controls support transparency reporting. Inventory completeness and use case categorization are agency responsibilities.
Risk management practices for AI	GOVERN, IDENTIFY	GOV-3, IDE-2	Partial	Severity classification (GOV-3) and impact assessment (IDE-2) produce attested risk categorization supporting evidence for NIST AI RMF alignment. Risk management practice design and implementation are agency responsibilities.
Rights-impacting and safety-impacting AI designation	IDENTIFY	IDE-2	Partial	Impact assessment attestation supports evidence for designation of rights-impacting and safety-impacting AI use cases with verifiable records. Designation decisions are made by agency officials.
Minimum risk management practices	GOVERN, MEASURE, RESPOND	GOV-1 through GOV-5, MEA-1 through MEA-4, RES-1 through RES-5	Partial	OVERT governance, measurement, and response domains collectively support evidence for minimum practices: impact assessment, monitoring, human oversight, and incident response — all with cryptographic attestation. Completeness and adequacy of minimum practices are agency responsibilities.
AI governance body oversight	GOVERN	GOV-2, HITL-4	Adjacent	Organizational role attestation (GOV-2) and policy approval attestation (HITL-4) document governance body decisions with tamper-evident records. Governance body establishment, composition, and oversight effectiveness are agency responsibilities.
Human oversight and appeal mechanisms	HITL	HITL-1 through HITL-4	Partial	OVERT human-in-the-loop controls provide attested consent (HITL-1), review (HITL-2), correction (HITL-3), and approval (HITL-4) workflows supporting evidence for human oversight requirements. Appeal mechanism design and due process obligations are agency responsibilities.

M-25-21 Requirement	OVERT Domain	OVERT Controls	Coverage	Alignment Notes
Ongoing monitoring and evaluation	MEASURE	MEA-1, MEA-2, MEA-4	Direct	S3P statistical safety signals provide continuous, quantitative, independently verifiable monitoring suitable for ongoing evaluation reporting.

M-25-22: AI in Federal Procurement

M-25-22 establishes requirements for the acquisition of AI capabilities by federal agencies, effective for solicitations issued on or after October 1, 2025 (180 days after issuance). OVERT controls support evidence for contractor compliance as follows:

M-25-22 Requirement	OVERT Domain	OVERT Controls	Coverage	Alignment Notes
AI performance and outcome tracking	MEASURE	MEA-1, MEA-2, MEA-4	Partial	S3P statistical safety signals provide quantitative performance metrics with confidence intervals within attested scope. OVERT tracks governance-control performance, not broad AI outcome tracking (accuracy, fairness, societal impact).
Data use restrictions and protections	PROTECT	PRO-1 through PRO-5	Partial	Boundary enforcement attestation supports evidence of data handling controls. Non-egress architecture (Section 17) supports evidence that declared boundary and non-egress controls executed within the attested scope. Data use restriction policy design is outside OVERT scope.
Transparency and explainability requirements	GOVERN	GOV-5, DISC-1	Partial	AI disclosure attestation provides verifiable transparency records. Machine-readable governance policy supports explainability documentation. Explainability of model decisions is outside OVERT scope.
Testing, evaluation, verification, and validation (TEVV)	MEASURE	MEA-3, MEA-4	Direct	OVERT TEVV controls produce attested test results at intervals defined in the operator's risk management policy.
Incident reporting and response	RESPOND	RES-1, RES-2, RES-5	Partial	Attested incident detection and response with cryptographic receipts. Declared failure modes (RES-5) support evidence for incident reporting requirements. Reporting obliga-

M-25-22 Requirement	OVERT Domain	OVERT Controls	Coverage	Alignment Notes
				tions and timelines are agency responsibilities.
Continuous monitoring	MEASURE, ATTEST	MEA-2, ATT-4	Direct	S3P and transparency log provide continuous monitoring infrastructure with independently verifiable records.
Vendor risk assessment	ATTEST	ATT-5	Adjacent	Notary network governance model attestation provides documented independent verification of vendor AI governance claims. Vendor risk assessment methodology and vendor selection are agency responsibilities.

Federal AI Procurement Alignment Summary

Agencies MAY reference OVERT conformance levels in Statements of Work (SOWs) or evaluation criteria as one mechanism for supporting evidence of AI governance practices:

- **OVERT Level 1 (Foundation):** Supports evidence for AI use case inventory and risk documentation requirements under M-25-21.
- **OVERT Level 2 (Enforcement):** Supports evidence for data use restriction and boundary enforcement requirements under M-25-22.
- **OVERT Level 3 (Measurement):** Supports evidence for continuous monitoring, TEVV, and human oversight requirements under both M-25-21 and M-25-22.
- **OVERT Level 4 (Evidence-Grade):** Provides the highest OVERT evidence and preservation tier for high-risk AI procurements, with governance records that are independently verifiable within the declared scope and subject to the denominator-source and scope disclosures required by Section 22.4.

OVERT conformance does not determine compliance with M-25-21, M-25-22, or any other federal requirement. Compliance determinations are made by contracting officers, agency Chief AI Officers, and other designated officials applying the requirements of each memorandum to specific acquisitions and use cases.

29. Crosswalk: Databricks AI Security Framework (DASF) v3.0

OVERT attestation artifacts support evidence for the mapped requirements below. Coverage qualifiers indicate the degree of alignment: **Direct** (OVERT directly produces the required artifact), **Partial** (OVERT supports some aspects but not all), **Adjacent** (OVERT evidence is relevant context). OVERT conformance does not determine compliance with DASF recommendations.

This crosswalk maps OVERT controls to the Databricks AI Security Framework (DASF) v3.0 compendium and companion agentic AI whitepaper. The compendium inventories 97 technical security risks across 13 AI system components and 73 mitigation controls. Relative to the older pre-agentic OVERT draft baseline, DASF v3.0 adds Component 13 (Agentic AI), 35 new agentic technical risks, and six new v3.0 controls (DASF 68-73); the compendium also includes controls DASF 65-67 introduced between the earlier v2.0 snapshot and the v3.0 agentic release. DASF is published by Databricks, Inc. under the Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0) license.

DASF describes what controls to configure across the AI system lifecycle. OVERT specifies how to produce cryptographic proof that runtime controls executed. The frameworks are complementary: DASF informs risk identification and control selection; OVERT provides the attestation layer that makes control execution independently verifiable. Where the companion PDF and earlier OVERT materials diverge on counts, this section follows the Databricks v3.0 compendium as the denominator source.

OVERT attestation is architecturally applicable to runtime enforcement, monitoring, governance, and agentic tool-use controls. DASF controls addressing training-time operations (data preparation, model training, experiment tracking), platform infrastructure (vulnerability management, SDLC, patching), and unmanaged client or secret-storage surfaces fall outside the OVERT attestation scope. These are noted as gaps below.

29.1 Risk-to-Control Crosswalk

The following table maps DASF risk categories to OVERT controls. Rows are grouped by DASF's 13 AI system components.

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
Data Operations (Components 1–4)				

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
1.1 Insufficient access controls	PROTECT, Agentic	CAP-1, PRO-5, TOOL-2	Direct	OVERT attests access enforcement at tool-call boundaries with per-action receipts; capability-based access control provides provenance-aware authorization
1.2 Missing data classification	IDENTIFY, GOVERN	IDE-1.2, GOV-3	Direct	System categorization and machine-readable risk taxonomy provide classification attestation artifacts
1.3 Poor data quality	MEASURE	MEA-2, MEA-4	Partial	S3P sampling detects quality degradation signals with confidence intervals; pre-deployment testing provides baseline. OVERT detects quality signals at runtime but does not address data quality management processes
1.4 Ineffective storage and encryption	PROTECT, ATTEST	PRO-2, ATT-1.4	Partial	OVERT attests network isolation state (NETATT) and TLS certificate pins; encryption implementation is outside OVERT scope
1.5 Lack of data versioning	—	—	—	No OVERT analog. Data versioning is an operational concern; OVERT attests policy and configuration versions
1.6 Insufficient data lineage	Agentic	CAP-1	Partial	Data provenance tracking covers lineage attestation for data flowing through tool calls; full-lifecycle data lineage is outside scope
1.7 Lack of data trustworthiness	ATTEST	ATT-1, ATT-4	Partial	Transparency log provides tamper-evident records of data access decisions with inclusion proofs; data quality and trustworthiness at source are outside scope
1.8 Legality of data	GOVERN, IDENTIFY	GOV-1, GOV-4, IDE-2	Adjacent	Governance policy and impact assessment support evidence of legal compliance documentation; supply chain governance covers third-party data. Legal compliance determination is outside OVERT scope
1.9 Stale data	—	—	—	No OVERT analog. Data freshness is an operational concern outside runtime attestation scope
1.10 Lack of data access logs	ATTEST, TOOL	ATT-1, ATT-4, TOOL-5	Direct	Tamper-evident, per-action audit trail with transparency log and notary attestation within attested scope

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
1.11 Compromised third-party datasets	GOVERN, PROTECT	GOV-4, PRO-1, PRO-4	Partial	Supply chain governance + boundary enforcement + input filtering with attestation receipts. OVERT attests enforcement at ingest boundaries; dataset integrity verification at source is outside scope
2.1 Preprocessing integrity	ATTEST	ATT-1, ATT-2	Partial	Co-epoch binding attests system configuration integrity; binary identity prevents unauthorized modification. Preprocessing logic correctness is outside scope
2.2 Feature manipulation	PROTECT	PRO-4, PRO-1	Partial	Input filtering + boundary enforcement attestation; runtime detection only
2.3 Raw data criteria	—	—	—	No OVERT analog. Data selection criteria are training-time decisions
2.4 Adversarial partitions	—	—	—	No OVERT analog. Train/test split manipulation is a training-time risk
3.1 Data poisoning	PROTECT, ATTEST	PRO-1, PRO-4, ATT-1	Partial	Boundary enforcement attests data ingestion policy compliance; transparency log records all decisions. Data poisoning prevention at the training level is outside scope
3.2 Ineffective storage and encryption (datasets)	PROTECT	PRO-2	Partial	Network isolation attestation (NETATT); storage encryption is infrastructure-level
3.3 Label flipping	—	—	—	No OVERT analog. Label manipulation is a training-time attack
4.1 Lack of traceability and transparency	GOVERN, ATTEST, Agentic	GOV-1, ATT-4, DISC-1.2	Direct	Machine-readable governance policy + transparency log + AI Bill of Materials
4.2 Lack of end-to-end ML lifecycle	GOVERN, HITL	GOV-1, GOV-2, HITL-4	Partial	Governance policy + accountability + configuration approval attestation; OVERT covers runtime governance within the ML lifecycle, not the full lifecycle
Model Operations (Components 5–8)				
5.1 Lack of experiment tracking	—	—	—	No OVERT analog. Experiment tracking is a development-time concern

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
5.2 Model drift	MEASURE, RESPOND, DRIFT	MEA-2, MEA-4, RES-1, DRIFT-2	Direct	S3P detects drift via violation rate changes with confidence intervals; adaptive control loop responds; DRIFT-2 detects behavioral drift within authorized bounds
5.3 Hyperparameters stealing	PROTECT	PRO-2, PRO-5	Partial	Network isolation + data isolation attestation supports evidence of parameter exfiltration prevention
5.4 Malicious libraries	GOVERN, ATTEST	GOV-4, ATT-2.2	Partial	Supply chain governance + binary identity attestation detects unauthorized code at attestation boundaries; library vetting is outside scope
6.1 Evaluation data poisoning	MEASURE	MEA-3, MEA-4	Partial	Third-party testing + pre-deployment testing provide independent evaluation; evaluation data integrity is outside scope
6.2 Insufficient evaluation data	MEASURE	MEA-3, MEA-4	Partial	OVERT mandates testing scope documentation published to transparency log; evaluation data sufficiency determination is outside scope
6.3 Lack of interpretability	GOVERN, Agentic	GOV-5, DISC-1	Adjacent	AI disclosure + transparency documentation; OVERT attests disclosure existence, not model interpretability
7.1 Backdoor/Trojaned model	ATTEST, GOVERN	ATT-2.2, GOV-4	Partial	Binary identity attestation detects model artifact tampering; supply chain governance. OVERT detects modification but cannot detect backdoors embedded before initial attestation
7.2 Model assets leak	PROTECT, ATTEST	PRO-2, PRO-5, ATT-1	Direct	Non-egress architecture + data isolation + network isolation prevent model exfiltration with attestation proof
7.3 ML supply chain vulnerabilities	GOVERN	GOV-4	Partial	Supply chain and third-party governance attestation; vulnerability assessment and remediation are outside scope
7.4 Source code control attack	—	—	—	No OVERT analog. Source code management is development infrastructure security
8.1 Model attribution	ATTEST, Agentic	ATT-1, ATT-4, DISC-1	Direct	Per-interaction attestation receipts provide cryptographic attribution; AI Bill of Materials

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
8.2 Model theft	PROTECT, ATTEST	PRO-2, PRO-5, ATT-1	Direct	Non-egress architecture prevents content exfiltration; attestation proves containment
8.3 Model lifecycle without HITL	HITL	HITL-1 through HITL-4	Partial	Runtime HITL attestation: consent, review, correction, and approval; full-lifecycle human oversight is broader than runtime attestation
8.4 Model inversion	PROTECT, ATTEST	PRO-1, PRO-3, ATT-1.2	Partial	Boundary enforcement + rate limiting + keyed commitments ensure content never leaves operator boundary; model inversion prevention effectiveness depends on rate limit configuration
Model Serving — Inference (Components 9–10)				
9.1 Prompt injection	PROTECT, TOOL	PRO-1, PRO-4, TOOL-1	Partial	Boundary enforcement + input filtering + pre-execution policy enforcement with attestation receipts; prompt injection prevention effectiveness depends on filter quality
9.2 Model inversion (serving)	PROTECT	PRO-3, PRO-1	Partial	Rate limiting + boundary enforcement attestation reduce query volume for extraction attacks
9.3 Model breakout	PROTECT, ATTEST	PRO-2, ATT-2	Direct	Network isolation attestation (NETATT) + co-epoch binding proves sandbox containment
9.4 Looped input	TOOL	TOOL-3.4, TOOL-3.3	Direct	Loop detection + circuit breaker with attested termination
9.5 Infer training data membership	PROTECT	PRO-3, PRO-1	Partial	Rate limiting attestation reduces query budget for membership inference; does not prevent membership inference attacks
9.6 Discover ML model ontology	PROTECT	PRO-3, PRO-1	Partial	Rate limiting + boundary enforcement attestation reduce model fingerprinting; does not prevent ontology discovery
9.7 Denial of service	PROTECT, TOOL	PRO-3, TOOL-3	Direct	Rate limiting + circuit breaking with attestation receipts
9.8 LLM hallucinations	MEASURE, PROTECT	MEA-2, PRO-4	Partial	S3P measures hallucination rates with confidence intervals; output filtering attestation. OVERT detects and measures halluci-

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
				nation rates but does not prevent hallucinations
9.9 Input resource control	PROTECT, TOOL	PRO-3, TOOL-3.1	Direct	Rate limiting + per-tool rate limits with attestation
9.10 Accidental data exposure	PROTECT, Agentic	PRO-5, CAP-1	Direct	Data isolation + capability-based access control with provenance tracking
9.11 Model inference API access	PROTECT, TOOL	PRO-1, TOOL-2	Direct	Boundary enforcement + function authorization with attested policy evaluation
9.12 LLM jailbreak	PROTECT, MEASURE	PRO-4, PRO-1, MEA-2	Partial	Input/output filtering + boundary enforcement + S3P violation rate monitoring; jailbreak prevention effectiveness depends on filter quality
9.13 Excessive agency	TOOL, Agentic, DRIFT	TOOL-1 through TOOL-5, CAP-1, CAP-2, DRIFT-3	Direct	Per-action policy enforcement with attestation receipts and capability scoping within attested scope; graph topology governance constrains agent proliferation. Prevention effectiveness depends on policy quality and scope completeness
10.1 Lack of inference quality monitoring	MEASURE, RESPOND	MEA-2, MEA-4, RES-1	Direct	S3P continuous measurement with confidence intervals + adaptive control loop
10.2 Output manipulation	PROTECT, ATTEST	PRO-4, ATT-1	Partial	Output filtering attestation + receipt integrity proves output was not tampered with post-attestation; pre-attestation output manipulation is outside scope
10.3 Discover model ontology (output)	PROTECT	PRO-3, PRO-4	Partial	Rate limiting + output filtering attestation reduce fingerprinting surface
10.4 Discover model family	PROTECT	PRO-3	Partial	Rate limiting attestation reduces fingerprinting attack surface
10.5 Black box attacks	PROTECT, MEASURE	PRO-3, PRO-4, MEA-2	Partial	Rate limiting + filtering + statistical monitoring with S3P; OVERT detects and rate-limits but does not prevent all black box attacks
10.6 Sensitive data output	PROTECT	PRO-4, PRO-5.3	Direct	Output filtering + PII detection attestation
Operations and Platform (Components 11–12)				

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
11.1 Lack of MLOps standards	GOVERN, HITL	GOV-1, GOV-2, HITL-4	Partial	Governance policy + accountability + configuration approval attestation for operational processes; MLOps standard definition is outside scope
12.1 Lack of vulnerability management	—	—	—	No OVERT analog. Infrastructure vulnerability management is outside OVERT scope
12.2 Lack of pen testing / red teaming	MEASURE	MEA-3	Partial	OVERT mandates third-party AI-specific testing; infrastructure penetration testing is outside scope
12.3 Lack of incident response	RESPOND	RES-1 through RES-5	Partial	Cryptographically gated incident response with attested escalation, override, revocation, and failure mode declaration within attested scope; IR team composition, training, and organization-wide incident response are outside scope
12.4 Unauthorized privileged access	HITL, ATTEST	HITL-4, ATT-5.3	Partial	Separation of duties attestation + notary independence requirements; access management systems are outside scope
12.5 Poor SDLC	—	—	—	No OVERT analog. Software development lifecycle is outside OVERT scope
12.6 Lack of compliance	GOVERN, IDENTIFY	GOV-1, GOV-3, IDE-2	Partial	Machine-readable governance policy + risk taxonomy + impact assessment support evidence for compliance; compliance determination is outside OVERT scope
12.7 Initial access	PROTECT	PRO-1, PRO-2	Partial	Boundary enforcement + network isolation attestation (NETATT); initial access prevention is primarily an infrastructure security concern
Agentic AI (Component 13)				
13.1 Memory poisoning	ATTEST, Agentic, PROTECT	ATT-1, ATT-4, CAP-1, PRO-5	Partial	OVERT can attest provenance of tool-fed context and enforce scoped retrieval within the attested boundary; poisoning of external memory stores or unmanaged prompt state remains outside scope
13.2 Tool misuse	TOOL, Agentic	TOOL-1 through	Direct	Pre-execution policy checks, per-call authorization, capability scoping, rate limits, and attestation receipts directly address unsafe

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
		TOOL-5, CAP-1, CAP-2		or unauthorized tool invocation within scope
13.3 Privilege compromise	Agentic, TOOL, PROTECT	CAP-1, CAP-2, TOOL-2, PRO-5	Direct	Capability-scoped identities, function authorization, and provenance-aware data isolation constrain privilege escalation and prove access decisions
13.4 Resource overload	PROTECT, RESPOND, TOOL	PRO-3, TOOL-3, RES-1	Direct	Rate limiting, circuit breaking, and adaptive responses are directly attestable
13.5 Cascading hallucination attacks	MEASURE, RESPOND, DRIFT, TOOL	MEA-2, RES-1, DRIFT-3, TOOL-3.4	Partial	OVERT can measure violation-rate shifts, detect recursive loops, and trip circuit breakers, but it cannot guarantee prevention of upstream hallucinations
13.6 Intent breaking and goal manipulation	PROTECT, TOOL, Agentic	PRO-1, PRO-4, TOOL-1, CAP-2	Partial	Boundary enforcement, input/output policy checks, and scoped capabilities reduce manipulation risk, but effectiveness depends on policy quality and mediation completeness
13.7 Misaligned and deceptive behaviors	MEASURE, Agentic, HITL	MEA-2, MEA-4, CAP-2, HITL-4	Partial	OVERT can surface anomalous behavior, require approvals, and constrain roles, but latent model misalignment is not fully solved by attestation
13.8 Repudiation and untraceability	ATTEST, TOOL	ATT-1, ATT-4, TOOL-5	Direct	Receipts, transparency logs, and attested action traces provide non-repudiation within scope
13.9 Identity spoofing and impersonation	HITL, ATTEST, TOOL	HITL-1.2, HITL-4.3, ATT-1, TOOL-2	Partial	OVERT can attest authenticated identities and approval chains in receipts, but identity provider controls remain external
13.10 Overwhelming human in the loop	HITL, RESPOND	HITL-1 through HITL-4, RES-1	Partial	OVERT can force approvals, escalations, and threshold-triggered intervention, but it cannot by itself optimize human workload or review quality
13.11 Unexpected RCE and code attacks	PROTECT, ATTEST, TOOL	PRO-2, ATT-2.2, TOOL-2	Direct	Isolation, binary identity attestation, and per-call authorization directly address code execution risk within attested execution boundaries

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
13.12 Agent communication poisoning	ATTEST, Agentic, PROTECT	ATT-1, ATT-4, MULTI-1, MULTI-2, PRO-2	Partial	Inter-agent trust boundaries, message attribution, and network isolation provide strong evidence within the attested graph, but semantic correctness of upstream agent outputs remains partly external
13.13 Rogue agents in multi-agent systems	Agentic, DRIFT, RESPOND	CAP-2, MULTI-1, MULTI-2, DRIFT-3, RES-4	Direct	Graph-topology governance, bounded delegation, revocation, and circuit breakers directly constrain rogue-agent proliferation within scope
13.14 Human attacks on multi-agent systems	HITL, Agentic, TOOL	HITL-4, CAP-2, MULTI-1, TOOL-1	Partial	Approvals, bounded delegation, and policy gating help, but operator misuse and social engineering remain partly external
13.15 Human manipulation	HITL, PROTECT, TOOL	HITL-1 through HITL-4, PRO-4, TOOL-1, CAP-2	Partial	Consent, review, correction, and policy gating reduce manipulation, but human susceptibility cannot be eliminated by attestation alone
13.16 Prompt injection (MCP server)	PROTECT, TOOL	PRO-1, PRO-4, TOOL-1, TOOL-2	Partial	Guardrails and pre-execution policy checks reduce unsafe tool use, but prevention depends on policy and filter quality
13.17 Confused deputy (MCP server)	TOOL, Agentic	TOOL-2, CAP-1, CAP-2	Direct	Function authorization and capability scoping directly address deputy misuse within scope
13.18 Tool poisoning (MCP server)	GOVERN, ATTEST, TOOL	GOV-4, ATT-2.2, TOOL-2, ATT-5	Partial	Supply-chain governance and binary identity attestation help detect poisoned tools, but upstream dependency vetting remains external
13.19 Credential and token exposure (MCP server)	PROTECT, ATTEST	PRO-2, PRO-5, ATT-1.4	Partial	Non-egress isolation, data containment, and transport-state evidence help, but secret storage and rotation controls are largely outside OVERT scope
13.20 Insecure server configuration (MCP server)	PROTECT, ATTEST, GOVERN	PRO-2, ATT-2.2, GOV-1	Partial	OVERT can attest configuration state at mediation boundaries and runtime identity, but full server hardening remains external

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
13.21 Supply chain attacks (MCP server)	GOVERN, ATTEST	GOV-4, ATT-2.2, ATT-5	Partial	Third-party governance and attested binary identity support evidence, but supply-chain risk management is broader than runtime attestation
13.22 Excessive permissions and scope creep (MCP server)	Agentic, TOOL	CAP-1, CAP-2, TOOL-2	Direct	Least-privilege capabilities and per-call authorization directly address scope creep within the attested boundary
13.23 Data exfiltration (MCP server)	PROTECT, ATTEST, TOOL	PRO-2, PRO-5, ATT-1, TOOL-2	Direct	Non-egress enforcement, data isolation, and per-call receipts directly support exfiltration prevention within scope
13.24 Context spoofing and manipulation (MCP server)	ATTEST, PROTECT, TOOL	ATT-1, ATT-4, PRO-4, TOOL-1	Partial	Provenance and policy checks help identify spoofed context, but upstream context trustworthiness is only partly attestable
13.25 Insecure communication (MCP server)	PROTECT, ATTEST	PRO-2, ATT-1.4	Partial	Network isolation and transport-state attestation support evidence of secure channels; protocol implementation details remain external
13.26 Malicious server connection (MCP client)	GOVERN, TOOL, ATTEST	GOV-4, TOOL-2, ATT-5	Partial	Third-party governance and attested connection policy support evidence, but complete vendor validation remains external
13.27 Insecure credential storage (MCP client)	PROTECT	PRO-2, PRO-5	Partial	OVERT constrains egress and data exposure within scope, but client credential storage itself is outside the attested runtime boundary
13.28 UI/UX deception (MCP client)	HITL, DISC, TOOL	HITL-1, HITL-2, HITL-3, DISC-1, TOOL-1	Partial	OVERT can attest disclosure, consent, and review checkpoints, not full interface-design integrity
13.29 Insufficient server validation (MCP client)	GOVERN, ATTEST, TOOL	GOV-4, ATT-5, TOOL-2	Partial	Attested trust-chain and external-service governance help, but complete server-validation logic is external
13.30 Client-side data leakage	PROTECT, Agentic	PRO-4, PRO-5, CAP-1	Partial	Output filtering, data isolation, and capability scoping reduce leakage within scope; client endpoint controls remain external
13.31 Excessive permission granting	Agentic, TOOL	CAP-1, CAP-2, TOOL-2	Direct	Capability scoping and per-call authorization directly address overbroad grants

DASF Component / Risk	OVERT Domain	OVERT Controls	Coverage	Notes
13.32 Client-side code execution	PROTECT, ATTEST	PRO-2, ATT-2.2, TOOL-2	Partial	Isolation and binary identity attestation help where client execution occurs inside the attested boundary; unmanaged client code remains external
13.33 Insecure communication handling	PROTECT, ATTEST	PRO-2, ATT-1.4	Partial	Network isolation and transport-state evidence help, but client transport-implementation details remain external
13.34 Session and state management failures	ATTEST, Agentic, HITL	ATT-1, ATT-4, CAP-1, HITL-4	Partial	Receipts, provenance, scoped capabilities, and approval chains help govern session state within scope; full session lifecycle controls are broader
13.35 Update and patch management	GOVERN, ATTEST	GOV-4, ATT-2.2	Partial	Runtime identity attestation and supply-chain governance support version-control evidence, but patch management is broader than OVERT

29.2 Control-to-Control Crosswalk

The following table maps DASF mitigation controls to OVERT controls, grouped by functional category.

DASF Control	Description	OVERT Controls	Coverage	Notes
Identity and Access				
DASF 1	SSO with IdP and MFA	HITL-1.2, HITL-4.3	Adjacent	OVERT attests authenticated identity in governance receipts; does not prescribe or implement authentication mechanism
DASF 2	Sync users and groups	GOV-2	Adjacent	Organizational role attestation; OVERT does not manage identity provisioning
DASF 5	Object-level access control	CAP-1, TOOL-2	Direct	Capability-based access control + function authorization with per-action attestation receipts
DASF 43	Access control lists	CAP-1, TOOL-2	Partial	Capability-based access + provenance-aware authorization; ACL management is outside scope

DASF Control	Description	OVERT Controls	Coverage	Notes
DASF 57	Attribute-based access control	CAP-1, TOOL-2	Partial	Provenance tracking supports evidence for attribute-based policy enforcement
DASF 64	Limit AI agent access	TOOL-1 through TOOL-5, CAP-1, CAP-2	Direct	Per-action policy enforcement + capability scoping + architectural separation within attested scope
DASF 67	Federate authentication	HITL-1.2, HITL-4.3, ATT-1, IDENT-1	Direct	IDENT-1 attests the full identity delegation chain including originating principal, each token exchange, scope narrowing verification, and token lifecycle. Federation protocol implementation remains external
Network and Isolation				
DASF 3	IP access lists	PRO-2	Partial	OVERT attests network isolation state (NETATT); IP list enforcement is infrastructure-level
DASF 4	Private link	PRO-2, ATT-1.4	Partial	Non-egress architecture + TLS certificate pinning attestation; private link configuration is infrastructure-level
DASF 34	Model isolation	PRO-2, PRO-5	Direct	Network isolation + data isolation attestation with co-epoch binding
DASF 56	Restrict outbound connections	PRO-2	Direct	NETATT attests egress policy, network controller identity, and eBPF state at each epoch
DASF 62	Network segmentation	PRO-2	Partial	Network isolation attestation covers segmentation enforcement at the AI boundary; broader network segmentation is infrastructure-level
MCP and Agent Hosting				
DASF 68	Use securely hosted managed MCP servers	GOV-4, TOOL-2, ATT-5, PRO-2, MCP-1	Direct	MCP-1 attests managed server identity, transport security, governance metadata, and per-call routing. Vendor internal operations remain external

DASF Control	Description	OVERT Controls	Coverage	Notes
DASF 69	Securely host custom MCP servers	GOV-4, ATT-2.2, PRO-2, TOOL-2, MCP-2	Direct	MCP-2 attests binary identity, network isolation, per-call authorization, and configuration change detection for operator-hosted MCP servers. Server deployment lifecycle remains external
DASF 70	Securely connect to external MCP servers	TOOL-2, PRO-2, ATT-5, CAP-1, MCP-3	Direct	MCP-3 attests connection governance, allowlist enforcement, capability scoping, output filtering, and connection lifecycle. External server internal posture remains external
DASF 72	Securely store and reuse agent state	PRO-5, ATT-1, ATT-4, CAP-1, STATE-1	Direct	STATE-1 attests state sealing, integrity verification, hash-chained lineage, mutation provenance, and access scoping. Storage infrastructure security remains external
DASF 73	Register prompts	GOV-1, DISC-1.2, ATT-4, TOOL-1, STATE-2	Direct	STATE-2 attests prompt registration, session binding, change detection, prompt-to-action traceability, and change approval governance. Prompt content quality and engineering methodology remain external
Data Security				
DASF 6	Classify data	IDE-1.2, GOV-3	Direct	System categorization + machine-readable risk taxonomy
DASF 8	Encrypt data at rest	—	—	No OVERT analog; encryption at rest is infrastructure-level
DASF 9	Encrypt data in transit	PRO-2	Partial	OVERT attests TLS certificate pins in NETATT; does not implement encryption
DASF 16	Secure model features	PRO-5, CAP-1	Partial	Data isolation + provenance-aware capability enforcement; feature store security is outside scope
DASF 51	Secure data sharing	PRO-5, ATT-1	Partial	Data isolation + non-egress attestation; data sharing governance is outside scope
DASF 46	Store and retrieve embeddings securely	PRO-5, PRO-2	Partial	Data isolation + network isolation attestation; embedding storage security is infrastructure-level

DASF Control	Description	OVERT Controls	Coverage	Notes
DASF 58	Data filters and masking	PRO-4, PRO-5.3	Direct	Output filtering + PII detection attestation
Audit and Monitoring				
DASF 14	Audit data actions	ATT-1, ATT-4, TOOL-5	Direct	Cryptographic upgrade: OVERT produces tamper-evident, notary-signed audit records vs. conventional logs
DASF 37	Inference tables	ATT-1, ATT-4, TOOL-5	Direct	Per-action attestation receipts in transparency log vs. mutable inference tables
DASF 55	Monitor audit logs	ATT-1, ATT-4, TOOL-5	Direct	OVERT audit trail is append-only with inclusion proofs and split-view detection
DASF 65	Implement end-to-end AI traceability	ATT-1, ATT-4, TOOL-5, DISC-1.2	Direct	Per-action receipts, transparency logs, and AI Bills of Materials provide end-to-end traceability within the attested scope
DASF 71	Log and register AI agents	ATT-1, ATT-4, DISC-1.2, GOV-1	Direct	Receipts, transparency logs, governance records, and AI Bills of Materials directly support agent inventory and activity logging within scope
DASF 21	Monitoring dashboard	MEA-2, RES-1	Adjacent	S3P provides quantitative monitoring signals; OVERT does not prescribe dashboard implementation
DASF 35	Track model performance	MEA-2, MEA-4	Partial	S3P statistical safety signals + pre-deployment testing requirements provide runtime performance measurement. DASF 35 encompasses broader model performance tracking including accuracy benchmarks, feature drift, and retraining triggers that extend beyond OVERT runtime attestation scope
DASF 36	Monitoring alerts	RES-1, RES-2	Direct	Adaptive control loop + incident response attestation
Guardrails and Enforcement				
DASF 31	Secure model serv-	PRO-1, PRO-2, PRO-3	Partial	Boundary enforcement + network isolation + rate limiting attestation for serving

DASF Control	Description	OVERT Controls	Coverage	Notes
	ing endpoints			endpoints; endpoint hardening is infrastructure-level
DASF 54	Implement AI guardrails	PRO-1, PRO-4	Direct	OVERT proves guardrails executed (not just configured): per-action permit/deny receipts with policy reference
DASF 60	Rate limiting	PRO-3, TOOL-3	Direct	Rate limiting with attested enforcement receipts and circuit breaking
Model Lifecycle				
DASF 18	Govern model assets	GOV-1, DISC-1.2	Partial	Governance policy attestation + AI Bill of Materials; model asset governance beyond runtime is outside scope
DASF 19	ML lifecycle management	GOV-1, GOV-2, HITL-4	Partial	Governance + accountability + approval attestation; covers runtime governance, not full ML lifecycle
DASF 23	Register/version/deploy model	HITL-4, GOV-1	Partial	OVERT attests deployment approvals; model registry is outside scope
DASF 24	Model access control	CAP-1, TOOL-2, PRO-5	Direct	Capability-based access + authorization + data isolation
DASF 29	MLOps workflows	GOV-1, HITL-4	Partial	Policy and approval attestation within operational workflows; MLOps workflow design is outside scope
DASF 42	MLOps/LLMOps	GOV-1, GOV-2	Adjacent	Governance attestation for ML operations processes; MLOps/LLMOps platform design is outside scope
Evaluation and Testing				
DASF 38	Pen testing and red teaming	MEA-3	Partial	OVERT mandates third-party AI testing; infrastructure pen testing is outside scope
DASF 45	Evaluate models	MEA-3, MEA-4	Direct	Third-party and pre-deployment testing requirements with transparency log publication
DASF 49	Automated LLM evaluation	MEA-2	Partial	S3P provides automated, statistically rigorous quality measurement for runtime behavior. DASF 49 encompasses broader au-

DASF Control	Description	OVERT Controls	Coverage	Notes
				tomated evaluation including pre-deployment benchmarks, evaluation dataset curation, and model comparison that extend beyond OVERT runtime attestation scope
Supply Chain and Compliance				
DASF 32	LLM provider management	GOV-4, ATT-5	Partial	Supply chain governance + notary governance for third-party verification; vendor management processes are outside scope
DASF 50	Platform compliance	GOV-1, GOV-3	Partial	Governance policy + risk taxonomy support evidence for compliance; compliance determination is outside OVERT scope
DASF 53	Third-party library control	GOV-4	Partial	Supply chain and third-party governance attestation; library vetting and vulnerability scanning are outside scope
Incident Response				
DASF 39	Incident response team	RES-1 through RES-5	Partial	Cryptographically gated incident response with attested escalation, override, revocation, and failure modes; IR team composition and training are outside scope
DASF 40	Internal access controls	HITL-4, ATT-5.3	Partial	Separation of duties + notary independence requirements; access control system implementation is outside scope
Human Oversight				
DASF 66	Use human-in-the-loop feedback	HITL-1 through HITL-4, RES-1	Partial	OVERT can attest consent, review, correction, approval, and escalation gates; broader feedback pipelines and product UX remain external
DASF 44	Event-triggered actions	RES-1	Direct	OVERT control loop triggers attested responses to threshold exceedances
DASF 48	Hardened ML runtime	ATT-2.2	Partial	Binary identity attestation proves runtime integrity via hardware-rooted measurement; runtime hardening implementation is outside scope
No OVERT Analog				

DASF Control	Description	OVERT Controls	Coverage	Notes
DASF 11	Capture and view data lineage	CAP-1	Partial	OVERT data provenance tracking covers tool-call data flows; full-lifecycle data lineage is outside scope
DASF 7, 10, 12, 13, 15, 17, 20, 22, 25, 26, 27, 28, 30, 33, 41, 47, 52, 59, 61, 63	Data quality checks, versioning, deletion, real-time data, EDA, training tracking, experiment tracking, representative data, RAG, fine-tuning, pre-training, model tags, encryption, secrets, secure SDLC, LLM comparison, source control, clean rooms, security training, software updates	—	—	These controls address training-time operations, data management, infrastructure security, or platform features outside the scope of runtime attestation. OVERT complements but does not replace these controls

29.3 Coverage Summary

DASF risks addressable by OVERT attestation:

- Section 29.1 intentionally distinguishes **Direct**, **Partial**, and **Adjacent** mappings. Those row-level qualifiers govern; aggregate percentages should not be read as equivalent to direct coverage or comprehensive DASF alignment.

- This revised crosswalk maps all **97** DASF v3.0 risk rows. **88 of 97** have at least a partial OVERT analog; **9 of 97** remain outside OVERT scope.
- All **35 Component 13 agentic risks** have at least a partial OVERT analog. The strongest alignment appears in runtime enforcement, tool-use governance, exfiltration controls, non-repudiation, circuit breaking, and multi-agent capability scoping.

DASF controls with OVERT analogs:

- Section 29.2 likewise distinguishes **Direct**, **Partial**, and **Adjacent** mappings. Combined mapping counts should not be interpreted as direct control equivalence.
- This revised crosswalk maps all **73** DASF controls. **58 of 73** have at least an adjacent or stronger OVERT analog; **15 of 73** remain outside OVERT scope.
- Six controls (DASF 67-70, 72-73) upgrade from Partial or Adjacent to Direct through the addition of MCP, STATE, and IDENT control families.

Key gaps — DASF risks not addressed by OVERT:

1. **Training-time attacks** (3.3 label flipping, 2.4 adversarial partitions, 5.1 experiment tracking): OVERT attests runtime behavior; training pipeline integrity is outside scope.
2. **Data lifecycle management** (1.5 versioning, 1.9 stale data, 2.3 raw data criteria): Operational data management concerns, not runtime governance.
3. **Platform infrastructure** (12.1 vulnerability management, 12.5 SDLC, 7.4 source code control): Traditional InfoSec controls outside AI attestation scope.
4. **Unmanaged secret and client surfaces** (13.19, 13.27, 13.30, 13.32-13.34): OVERT can attest bounded behavior within the mediated runtime, but credential storage, unmanaged clients, and full session lifecycle controls remain partly external.

Key gaps — OVERT controls not addressed by DASF:

1. **Cryptographic attestation** (ATT-1 through ATT-5): DASF has no equivalent to non-egress attestation, co-epoch binding, three-phase attestation, transparency logs, or notary governance. This is OVERT's primary contribution.
2. **Statistical safety measurement** (MEA-1, MEA-2): DASF recommends monitoring (DASF 21, 35, 36) but does not specify cryptographically verifiable, auditor-reproducible measurement with exact confidence intervals.
3. **Risk signals** (OVERT risk signals (see Section 4.6 and Annex D)): DASF has no equivalent verifier-usable signal architecture. OVERT provides quantitative, independently verifiable runtime signals within the declared mediation scope for monitoring, audit, and external risk analysis.
4. **Evidence-grade agentic runtime control** (TOOL-1 through TOOL-5, CAP-1/CAP-2, MULTI-1/MULTI-2): DASF v3.0 adds meaningful agentic controls (64-73), especially around

MCP, agent registration, state, and prompt governance; OVERT still provides per-action attestation, capability-based access control, inter-agent trust boundaries, and loop detection with independent verification of execution.

5. **Failure mode and override** (RES-3, RES-4, RES-5): DASF 39 covers incident response broadly; OVERT specifies attested emergency override, scoped revocation, and explicit fail-open/fail-closed declarations.

Differentiation summary: DASF v3.0 materially narrows the conceptual gap by explicitly modeling agents, MCP servers, MCP clients, agent state, and prompt governance. OVERT remains differentiated because it upgrades those runtime controls within the attested scope from configuration guidance to tamper-evident, independently assessable execution evidence. Organizations deploying in regulated environments can use DASF for risk identification and control selection, then use OVERT to prove that the most consequential runtime controls actually executed.

29.4 Attestation Boundary Declaration

Conformant implementations SHALL publish an **Attestation Boundary Declaration** (ABD) specifying which system surfaces are within the attested runtime boundary and which are outside it. The ABD SHALL be published to the transparency log and referenced in the conformance statement.

The ABD SHALL address, at minimum:

Surface Category	In-Scope Indicator	Out-of-Scope Indicator
MCP servers (managed)	Server identity, transport, governance metadata attested per MCP-1	Vendor internal operations, hosting lifecycle
MCP servers (custom)	Binary identity, network isolation, authorization attested per MCP-2	Deployment automation, patching lifecycle
MCP servers (external)	Connection governance, capability scoping, output filtering attested per MCP-3	External server internal security posture
Agent durable state	State sealing, lineage, mutation provenance attested per STATE-1	Storage infrastructure security, backup/recovery
Prompt artifacts	Registration, binding, change governance attested per STATE-2	Prompt engineering methodology, content quality
Identity delegation	Delegation chain, scope narrowing, token lifecycle attested per IDENT-1	IdP implementation, credential storage
Unmanaged clients	(Not attestable — outside runtime boundary)	Client-side code, credential storage, UI integrity

Surface Category	In-Scope Indicator	Out-of-Scope Indicator
Secret storage	(Not attestable — outside runtime boundary)	Secret rotation, vault implementation, access control

Where a surface is partially in scope (e.g., the arbiter attests transport security to an external MCP server but not the server's internal posture), the ABD SHALL state the boundary precisely.

Annex A: Glossary (Informative)

The following terms and acronyms are used throughout this standard. Where a term has a specific OVERT definition that differs from common usage, the OVERT-specific definition is provided.

Term ID	Term	Definition
A.1	AAL	**Attestation Assurance Level.** One of four tiers (AAL-1 through AAL-4) describing the cryptographic verifiability and independence of governance attestation artifacts. AAL-1: Policy Documentation (self-asserted). AAL-2: Process Records (self-attested, auditor must trust operator). AAL-3: Automated Monitoring (machine-generated, operator-controlled). AAL-4: Cryptographic Attestation (third-party verifiable, zero content access required). See Section 4.1.
A.2	Arbiter	The enforcement sidecar that intercepts AI system actions (tool calls, API requests, data access) and evaluates them against policy before permitting execution. Implemented as an enforcement module; the specific runtime technology is specified by the applicable Protocol Profile. The arbiter generates attestation envelopes for every intercepted action.
A.3	OVERT	**Observable Verification Evidence for Runtime Trust.** This standard.
A.4	Attestation Artifact	A cryptographically signed record produced by the OVERT attestation infrastructure demonstrating that a specific governance control executed at a specific time under a specific configuration. Includes envelopes, receipts, S3P attestations, and ControlActions.
A.5	Attestation Pack	A bundled collection of attestation artifacts sufficient to demonstrate conformance for a defined scope and time period. Includes receipts, transparency log proofs, epoch data, S3P attestations, and ControlActions.
A.6	BLS	**Boneh-Lynn-Shacham.** A pairing-based signature scheme permitting efficient aggregation of multiple signers' signatures into a single compact signature. Used for notary threshold signatures in Protocol Profile 1.0. Not post-quantum resistant; the standard requires hybrid classical + post-quantum constructions after January 1, 2031. Alternative notary signature constructions (e.g., multi-signature with Ed25519 or ML-DSA) may be specified by other Protocol Profiles.
A.7	CAS	**Content-Addressable Storage.** Local storage within the operator's environment where attestation evidence (prompts, responses, evaluations) is stored indexed by cryptographic commitment. Content never leaves the operator's boundary.

Term ID	Term	Definition
A.8	CBOR	**Concise Binary Object Representation.** Binary data serialization format (RFC 8949). Protocol Profile 1.0 uses CBOR deterministic encoding per Section 4.2 of RFC 8949 for canonical byte-level representation. The standard requires deterministic encoding as a property (Section 17.1); the specific format — CBOR, JSON via JCS (RFC 8785), or other deterministic encoding — is specified by the applicable Protocol Profile.
A.9	CI (Confidence Interval)	A statistical interval computed using the Clopper-Pearson exact method providing upper and lower bounds on a violation rate with guaranteed coverage probability. Used in S3P attestations.
A.10	Co-epoch Binding	The cryptographic binding of an attestation receipt to the binary identity, network isolation state, and configuration of the system under attestation at the time the attestation was produced. Ensures that attestations cannot be replayed across different system configurations.
A.11	ControlAction	A cryptographically attested record of a governance response to a detected violation. Includes action type, timestamp, scope, and co-epoch binding.
A.12	DPL	**Digest Publication Ledger.** A per-epoch publication of request commitments (never raw digests) enabling auditor verification of sampling fairness without content access.
A.13	Epoch	A bounded time interval (configurable; recommended default: 300 seconds) during which attestation parameters remain constant. Epoch boundaries trigger nonce publication, key rotation, and S3P computation.
A.14	HKDF	**HMAC-based Key Derivation Function.** Key derivation per RFC 5869. Protocol Profile 1.0 uses HKDF for deriving tenant_pepper, storage_key, sampling_key, and epoch_secret from root secrets within the split-knowledge key hierarchy. The standard requires key derivation as specified by the applicable Protocol Profile.
A.15	HMAC	**Hash-based Message Authentication Code.** Keyed hash function per RFC 2104. Protocol Profile 1.0 uses HMAC for request commitments, evidence commitments, PRF tags, and S3P sampling tags with domain separation prefixes. The standard requires keyed commitment functions as specified by the applicable Protocol Profile.
A.16	IAP	**Independent Attestation Provider.** An entity structurally independent of the AI system operator that operates notary infrastructure, validates attestations, and publishes transparency log entries. An IAP does not access protected content. Multiple IAPs may operate under different governance models.
A.17	Base Envelope	The baseline attestation envelope emitted for every AI request. Contains 9 fields including blinded identifier, request commitment, encoder binary identity, and

Term ID	Term	Definition
		metadata. Closed schema — no additional fields permitted. See Annex B for architecture; full field-level schema is in the Protocol Profile.
A.18	Extended Envelope	The extended attestation envelope emitted for sampled requests. Contains 10 fields including the full PRF tag for auditor recomputation and policy evaluation scores. Closed schema. See Annex B for architecture; full field-level schema is in the Protocol Profile.
A.19	NETATT	**Network Attestation.** A per-epoch attestation of the system's network isolation state, including egress policy, network policy controller identity, eBPF state, CNI configuration, environment variables, and TLS certificate pins. Bound to all receipts issued during the epoch.
A.20	OSCAL	**Open Security Controls Assessment Language.** NIST-developed machine-readable format for security control documentation. OVERT attestation packs are expressible as OSCAL Assessment Results.
A.21	POST_HOC Receipt	An attestation receipt generated retroactively after a fail-open period (per RES-5.2). POST_HOC receipts are reconstruction artifacts and SHALL NOT be counted as contemporaneous attestation coverage for conformance, risk-signal reporting, or litigation reporting purposes. Distinguishable from contemporaneous receipts in all export packages and signal computations.
A.22	PRF	**Pseudorandom Function.** A deterministic function (HMAC-SHA256 in Protocol Profile 1.0) used to determine whether a given request falls within the attestation sample. Operates on request commitments, not raw content.
A.23	Protocol Profile	A registered implementation specification defining cryptographic constructions, envelope schemas, key derivation methods, and receipt formats that implement this standard. Multiple profiles may coexist. Conformance requires exactly one registered profile per deployment. See Annex B for Protocol Profile 1.0 summary.
A.24	RATS	**Remote ATtestation procedureS.** IETF architecture for remote attestation (RFC 9334). OVERT attestation architecture is complementary to RATS, with roles mapping to Attester (arbiter), Verifier (notary), and Relying Party (auditor/insurer).
A.25	Receipt	A cryptographically signed record issued by the notary service proving that a specific attestation envelope was received, validated, and recorded during a specific epoch. Contains 9 fields including attestation_hash, epoch binding, binary identity, network state, flags (contemporaneous vs. POST_HOC), and transparency log proofs. See Annex B for architecture; full field-level schema is in the Protocol Profile.
A.26	S3P	**Statistical Safety Signal Protocol.** The normative auditor-reproducible sampling and measurement method defined in Section 9 (MEASURE). Uses

Term ID	Term	Definition
		commit-then-reveal epoch nonces, keyed-function-based sampling (HMAC in Protocol Profile 1.0), and Clopper-Pearson exact confidence intervals to produce statistically rigorous safety signals.
A.27	Severity Class	A classification of governance violations by severity, defined in GOV-3.2. Maps to risk-signal computation and response escalation requirements.
A.28	Split-Knowledge Key Hierarchy	The key management architecture in which content-binding keys (operator-managed, e.g., tenant_pepper) and sampling/identity keys (platform-managed, e.g., sampling_key, epoch_secret) are held by different parties. Ensures that routine audit is a zero-content-knowledge operation.
A.29	SPKI	**Subject Public Key Info.** DER-encoded public key information used for TLS certificate pinning in NETATT.
A.30	STH	**Signed Tree Head.** A signed commitment to the current state of the transparency log Merkle tree, enabling split-view detection.
A.31	SUT	**System Under Test.** The AI system being governed. The attestation system treats the SUT as untrusted — self-reports from the SUT are insufficient for AAL-4 conformance. This designation is specific to the OVERT attestation relationship and is distinct from the NIST SP 800-207 Zero Trust Architecture for network security.
A.32	TEVV	**Test, Evaluation, Verification, and Validation.** The systematic process of evaluating AI system performance, safety, and governance compliance. OVERT provides the attestation infrastructure for TEVV activities.
A.33	Transparency Log	An append-only, cryptographically verifiable log (per RFC 6962) in which attestation receipts and S3P attestations are recorded. Supports inclusion proofs (proving a receipt is in the log), consistency proofs (proving the log has not been tampered with), and signed tree heads for split-view detection.
A.34	Notary Network	One or more notary nodes operated under a published governance model (ATT-5.1) that validate attestations on behalf of relying parties. Where multiple nodes are deployed, t-of-n agreement is required before a valid receipt can be issued and no single node can unilaterally issue or suppress a receipt. A single structurally independent notary satisfies the AAL-4 independence requirement but not the AAL-4 resilience requirement; single-IAP deployments SHALL disclose this limitation per Section 4.7.1 and the conformance statement grammar (Section 22.4). Multi-entity sets provide both independence and resilience. The signature or verification construction achieving the t-of-n property (where applicable) is specified in the registered Protocol Profile.

Annex B: Protocol Profile Reference Summary

> *Protocol Profile 1.0 is the initial registered profile authored by GLACIS Technologies. Additional profiles may be submitted by any party meeting the registration requirements defined in Section 22.6. This annex summarizes Protocol Profile 1.0 for reference. The authoritative specification is the Protocol Profile document itself.*

B.1 Cryptographic Primitives

Protocol Profile 1.0 specifies the following cryptographic constructions:

Primitive	Usage	Specification
SHA-256	All digests, binary hashes, commitments	FIPS 180-4
HMAC-SHA256	PRF tags, request/evidence commitments, bearer tokens, S3P sampling	RFC 2104
HKDF-SHA256	Key derivation for both operator and platform key hierarchies	RFC 5869
Ed25519	Arbiter signatures, notary signatures, controller signatures	RFC 8032
Notary signature (BLS threshold in Profile 1.0)	Notary network t-of-n verification	draft-irtf-cfrg-bls-signature (construction selected by Protocol Profile 1.0; the core standard requires only the t-of-n trust property and permits alternative constructions including multi-signature schemes)
Deterministic encoding (CBOR in Profile 1.0, JCS for JSON profiles)	Canonical encoding of attestation structures	RFC 8949, Section 4.2 (CBOR); RFC 8785 (JCS)

Primitive	Usage	Specification
Merkle trees	Transparency log inclusion and consistency proofs	RFC 6962
Clopper-Pearson	Exact binomial confidence intervals for S3P	Standard statistical method

Post-quantum migration path: Protocol Profile 1.0 uses BLS threshold signatures and Ed25519 single-signer signatures, both of which rely on the computational Diffie-Hellman problem and are vulnerable to quantum attack via Shor's algorithm. Section 18 requires that after January 1, 2031, conformant implementations use hybrid classical + post-quantum constructions. For single-signer operations, the recommended migration is ML-DSA (FIPS 204) + Ed25519. For notary network operations, Protocol Profiles using multi-signature constructions can migrate each notary independently to ML-DSA; profiles using threshold signatures require threshold-compatible post-quantum schemes. Pure classical signature schemes become non-conformant after that date, per NIST IR 8547 deprecation timeline.

IETF RATS alignment and EAT forward reference: OVERT attestation envelopes are structurally aligned with the IETF RATS architecture (RFC 9334). The Entity Attestation Token (EAT, RFC 9711) defines a CBOR/JSON token format for attester-generated claims about entity identity and state. Future Protocol Profile revisions should evaluate expressing OVERT attestation envelopes as EAT profiles, enabling interoperability with the broader RATS ecosystem. In particular, an Agentic AI EAT Capability Attestation profile — encoding arbiter binary identity, co-epoch binding, and capability-scoped policy claims as EAT claims — would position OVERT attestation artifacts for direct consumption by EAT-aware verifiers and relying parties within the IETF trust model. This alignment is a design goal for Protocol Profile 2.0 and does not affect Protocol Profile 1.0 conformance.

B.2 Domain Separation and Key Architecture

Protocol Profile 1.0 uses versioned domain separation prefixes on all HMAC operations to prevent cross-protocol attacks. Each HMAC operation — request commitment, evidence commitment, sampling PRF, epoch bearer token, and S3P sampling — uses a distinct prefix with a version suffix enabling future protocol evolution.

The split-knowledge key hierarchy ensures that content-binding keys (operator-managed) and sampling/identity keys (platform-managed) are held by different parties. This separation prevents any single party from both reversing content AND verifying sampling fairness. The specific prefix strings, salt values, and key derivation parameters are specified in the Protocol Profile.

B.3 Canonicalization

Protocol Profile 1.0 canonicalizes all OVERT messages per RFC 8949 Section 4.2 (Deterministic Encoding). Two conformant encoders encoding the same logical data must produce identical byte sequences, which is the prerequisite for all hash-based verification. The standard requires deterministic canonicalization as a property (Section 17.1); the specific encoding format is specified by the applicable Protocol Profile. Protocol Profiles using JSON are expected to specify JCS (RFC 8785) for deterministic canonicalization; Protocol Profiles using CBOR are expected to specify RFC 8949 Section 4.2.

Protocol Profile 1.0 prohibits IEEE-754 floating-point numbers in attestation envelopes, requiring scaled integers for deterministic cross-platform hashing. The S3P schema uses decimal strings for rates and bounds. Timestamps use uint64 nanoseconds since Unix epoch. Indefinite-length encodings, NaN, and +/-Inf are rejected. Protocol Profiles using JSON encodings are expected to specify equivalent numeric safety requirements (e.g., string-encoded decimals, integer-only numeric fields, or explicit precision annotations) to satisfy the numeric losslessness property of Section 17.1.1.

B.4 Commitment Architecture

Protocol Profile 1.0 defines a layered commitment architecture:

- **Request commitments** are computed by HMAC over the content digest using an operator-managed key derived from the operator's root secret via HKDF. The content digest stays local; only the HMAC commitment crosses the trust boundary.
- **Evidence commitments** follow the same pattern for policy evaluation evidence.
- **PRF sampling tags** are computed using a platform-managed key, operating on the request commitment (not the raw content digest). This ensures auditors can verify sampling fairness without holding content-reversing keys.

The specific HMAC constructions, HKDF derivation parameters, and key hierarchy are specified in the Protocol Profile.

Critical constraint: Operator-managed content-binding keys never leave the operator's environment. Deep audits requiring content verification are conducted on-premises under operator control and legal authority.

B.5 Key Hierarchy

The split-knowledge key hierarchy has two branches:

Operator-managed keys (content binding): Derived from an HSM-backed root secret. Includes keys for content commitment and local storage encryption. These keys never cross the operator's trust boundary.

Platform-managed keys (identity and sampling): Derived from an HSM-backed root secret managed by the notary network operator. Includes keys for sampling, epoch management, and notary signing. In Protocol Profile 1.0, notary signing keys are BLS threshold shares distributed across notary nodes; other Protocol Profiles may use per-notary signing keys with independent key management.

Forward secrecy: Epoch-scoped keys are deleted after the subsequent epoch begins. Compromise of a current epoch secret does not reveal past epoch secrets.

Recovery: Shamir k-of-n (recommended: 3-of-5) across geographic regions or HSMs for operator root secrets. Platform key recovery is specified in the registered Protocol Profile (Protocol Profile 1.0 uses BLS threshold key shares across notary nodes; multi-signature profiles use standard per-node key backup procedures).

The specific HKDF derivation paths, salt values, and key tree structure are specified in the Protocol Profile.

B.6 Attestation Envelope Architecture

Protocol Profile 1.0 defines three closed-schema structures:

Base Envelope (all requests — 9 fields): Emitted for every in-scope AI action. Contains a blinded identifier, request commitment, encoder binary identity, non-content metadata, monotonic counter, nanosecond timestamp, key identifier, arbiter instance identifier, and signature. No additional fields are permitted (closed schema).

Extended Envelope (sampled requests — 10 fields): Emitted for requests selected by the sampling PRF. Contains a reference to the matching Base Envelope, request and evidence commitments, the full PRF tag for auditor recomputation, policy evaluation scores, monotonic counter, timestamp, key and arbiter identifiers, and signature. Closed schema.

Receipt (9 fields, issued by notary service): Contains the attestation hash (cryptographic digest of the submitted envelope), validated epoch, notary-derived binary hash, network state hash, monotonic counter, issuance timestamp, flags (contemporaneous vs. POST_HOC), notary signature

(single-signer, multi-signature, or threshold, as specified in the Protocol Profile), and transparency log proofs (inclusion proof, consistency proof, signed tree heads). Closed schema.

The `flags` field distinguishes contemporaneous receipts (`0x00`) from POST_HOC receipts (`0x01` , generated after a fail-open period per RES-5.2). Auditors and risk-signal computations filter on this field to separate contemporaneous attestation from retroactive reconstruction.

The complete field-by-field schemas with types, constraints, and signature scopes are specified in the Protocol Profile.

B.7 S3P Attestation Schema

The S3P attestation schema is a 14-field closed structure capturing all data needed for auditor-reproducible safety verification. Every field is necessary and sufficient for independent recomputation of statistical bounds.

The schema includes: epoch identifier, violation type, total and sampled request counts, sampling and observed rates (decimal strings to avoid IEEE-754 variance), confidence level, Clopper-Pearson lower and upper bounds (decimal strings), sampling threshold, epoch nonce commitment, status indicator, and notary signature.

The three status values are: `"OK"` (valid computation), `"ERR_INSUFFICIENT_SAMPLE"` (sample size below minimum), and `"ERR_NONCE_NOT_PUBLISHED"` (verification failure — epoch nonce was not published after epoch close).

The complete schema with field types and encoding rules is specified in the Protocol Profile.

B.8 Clopper-Pearson Confidence Interval Computation

The Clopper-Pearson method provides exact (not approximate) binomial confidence intervals with guaranteed coverage probability. It provides exact binomial interval coverage under the S3P sampling model and remains valid for small sample sizes without normal-approximation assumptions. The upper bound is conservative by construction.

Given `k` violations observed in `n` sampled requests at confidence level `1 - alpha` :

```
CI_lower = Beta_inv(alpha/2;     k,     n - k + 1)     for k > 0, else 0
CI_upper = Beta_inv(1 - alpha/2; k + 1, n - k)         for k < n, else 1
```

Where `Beta_inv(p; a, b)` denotes the p-th quantile of the Beta distribution with shape parameters `a` and `b`.

Properties:

- Exact coverage: P(p_true in [CI_lower, CI_upper]) >= 1 - alpha for all p_true
- Conservative: The interval is wider than approximate methods (Wald, Wilson), never narrower
- Valid for all sample sizes, including small samples
- No distributional assumptions required

B.9 Receipt Service Architecture

The receipt service accepts a closed-schema request containing only a hash and an epoch identifier, and returns a signed receipt. The API schema enforces the non-egress architecture at the protocol level: the service is structurally incapable of receiving content because its schema does not contain fields for content. Unknown fields are rejected.

This constraint is architectural, not merely a validation rule. The receipt service API schema is specified in the Protocol Profile.

B.10 Informative Latency Targets

The following latency targets are informative recommendations for Protocol Profile 1.0. Specific latency requirements are deployment-dependent and are not normative requirements of the standard.

Phase	Operation	Informative Target
Phase 1 — Enforcement	Local policy evaluation	< 5 ms P50
Phase 1 — Enforcement	Distributed policy evaluation	< 25 ms P50
Phase 2 — Attestation	Receipt round-trip	< 50 ms P50
Phase 3 — Commitment	Transparency log inclusion	< 100 ms P95

Total overhead (enforcement + attestation): informative target < 50 ms P50, which is negligible relative to typical LLM inference latency (500-5000 ms).

B.11 Informative Default Parameters

The following default parameters are informative recommendations for Protocol Profile 1.0. Operators configure these values according to their deployment requirements.

Parameter	Informative Default	Notes
Epoch duration	300 seconds (5 minutes)	Configurable per deployment policy
Tool-call recursion depth	25	Configurable threshold defined in deployment policy
Clock skew tolerance	<= 2 seconds	Bounded skew tolerance; stale submissions rejected
Override review SLA	Within operator-defined SLA	Recommended: 24 hours
TEVV testing interval	Per operator's risk management policy	Not to exceed 12 months or as required by applicable regulation

B.12 Implementation Resources

Protocol Profile 1.0 includes CBOR diagnostic notation examples, reference test vectors for S3P computation, and auditor verification procedures. These materials enable implementers to validate their implementations against known-good results. Protocol Profiles using other encodings are expected to provide equivalent notation examples and test vectors in their respective formats.

Reference test vectors and implementation examples are available in the Protocol Profile document. Organizations implementing OVERT using Protocol Profile 1.0 should obtain the Protocol Profile from the OVERT Protocol Profile Registry.

Annex C: Design Rationale and Case Studies

> *This annex provides design rationale and contextual analysis. It describes legal, operational, and institutional conditions relevant to the standard's development. It does not impose requirements on implementers or assert legal conclusions. The normative requirements of the standard are specified in Parts 1-5. It is structured as Design Decision, Rationale, and Supporting Analysis.*

C.1 Verification Gaps in High-Stakes AI Deployments

Design Decision: OVERT requires independent, third-party verifiable attestation (AAL-4) for governance controls in high-stakes deployments, rather than relying on self-attestation or contractual governance alone.

Rationale: Contractual governance has proven structurally insufficient as the sole enforcement mechanism for AI safety controls. When disputes arise between AI system providers and their customers over safety control execution, neither party can independently verify what controls actually ran if no attestation infrastructure exists.

Supporting Analysis: In early 2026, a series of disputes between major AI laboratories and government agencies demonstrated this proof gap with extraordinary clarity. In one instance, an AI company insisted on contractual red lines regarding prohibited uses, while the government customer demanded unrestricted operational access. Neither party could independently verify whether AI use complied with stated restrictions during operational deployment. The dispute was adjudicated through contract negotiations, leaked internal memoranda, public conference statements, and executive action — rather than through independent verification of actual system behavior.

Simultaneously, competing AI laboratories publicly accused each other of inadequate safety practices, with characterizations ranging from "safety theater" to "mendacious" claims about governance controls. These mutual accusations could not be independently adjudicated because no party had deployed infrastructure capable of producing verifiable records of what safety controls actually executed on any given interaction. The disputes were resolved — or remain unresolved — through political, commercial, and reputational channels rather than through technical verification.

This pattern illustrates a structural problem: contractual governance produces assertions about intended behavior, not verifiable records of actual behavior. When the only evidence of safety control execution is the operator's own claims, disputes become contests of credibility rather than questions of fact. If verification technology becomes commercially deployable at scale, the continued reliance on unverifiable self-attestation may become relevant to courts, regulators, and insurers evaluating evidentiary and governance posture under applicable legal and policy frameworks.

OVERT addresses this gap by specifying how to produce tamper-evident, independently verifiable, temporally bound proof that AI governance controls executed — without exposing protected content.

C.2 The T.J. Hooper Principle and Potential Standard-of-Care Analysis

Design Decision: OVERT is designed as an open standard that can serve as one reference point in discussions of verifiable AI governance, recognizing that courts — not industries — ultimately determine the standard of care.

Rationale: The T.J. Hooper principle holds that an entire industry can be found negligent for failing to adopt available safety technology, regardless of industry custom.

Supporting Analysis: In The T.J. Hooper, 60 F.2d 737 (2d Cir. 1932), Judge Learned Hand held that tugboat operators were negligent for failing to carry radio receivers that would have warned of an approaching storm — even though no tugboat company used radios at the time. The court stated: "a whole calling may have unduly lagged in the adoption of new and available devices… Courts must in the end say what is required; there are precautions so imperative that even their universal disregard will not excuse their omission."

The principle has been applied consistently for nearly a century. The RAND Corporation's report on AI tort liability explicitly cited T.J. Hooper, noting that "courts can still find AI companies negligent even if they did follow industry custom" and that safety-conscious companies developing standards "could establish benchmarks for the whole industry in future litigation."

For AI governance, the implication is narrower. If cryptographic attestation technology becomes commercially deployable and operationally mature, failure to consider its adoption could become relevant to negligence analysis, depending on jurisdiction, commercial availability, deployment maturity, and the surrounding facts. An open standard may strengthen that analysis by documenting an interoperable approach, but publication of a standard alone does not establish that the technology is available, required, or legally obligatory.

The English equivalent is Bolitho v. City and Hackney Health Authority [1998] AC 232, where Lord Browne-Wilkinson held that courts may reject professional custom as unreasonable if "the professional opinion is not capable of withstanding logical analysis." The Australian statutory calculus under Section 5B of the Civil Liability Act 2002 (NSW) reaches the same outcome through explicit consideration of the probability of harm, seriousness of harm, burden of precautions, and social utility. The German doctrine of Verkehrssicherungspflichten requires anyone who creates or controls a potential source of danger to take necessary precautions.

C.3 Adverse Inference Doctrine and the Duty to Create Records

Design Decision: OVERT Section 21 (Legal Preservation and Production) requires retention policies, legal hold procedures, immutable export capabilities, and chain-of-custody metadata — addressing the risk that operators could "define away bad evidence" through self-serving retention policies.

Rationale: The adverse inference doctrine permits factfinders to draw unfavorable conclusions when a party fails to preserve records it had the available technology to produce. For AI systems, where the "black box" nature makes governance documentation critical, the absence of attestation technology may be relevant to evidentiary analysis. Whether the failure to deploy attestation creates a substantive claim or an evidentiary disadvantage depends on jurisdiction, applicable duty, commercial availability of attestation technology, and the specific facts. This rationale identifies the doctrinal relevance; it does not assert a specific legal outcome.

Supporting Analysis: Under FRCP 37(e), if electronically stored information that should have been preserved is lost because a party failed to take reasonable steps to preserve it, the court may order measures no greater than necessary to cure the prejudice. Upon finding that a party acted with the intent to deprive another party of the information's use in the litigation, the court may presume that the lost information was unfavorable to the party, instruct the jury that it may or must presume the information was unfavorable, or dismiss the action or enter a default judgment.

In Zubulake v. UBS Warburg, 229 F.R.D. 422 (S.D.N.Y. 2004), failure to preserve digital evidence resulted in a $29.2 million verdict including $21.1 million in punitive damages. The court granted an adverse inference instruction: "if you find that UBS could have produced this evidence… you are permitted, but not required, to infer that the evidence would have been unfavorable to UBS."

Valcin v. Public Health Trust of Dade County, 473 So.2d 1297 (Fla. 3d DCA 1984), provides the closest doctrinal analogue. Where a hospital's file failed to contain an operative note, the court imposed a rebuttable presumption of negligence and shifted the burden of proof for records that should have been created pursuant to a duty. If industry standards (NIST AI RMF, ISO/IEC 42001)

and available technology create a de facto duty to record AI safety control execution, the failure to deploy attestation technology triggers a parallel adverse presumption.

OVERT Section 21 directly mitigates this risk by requiring operators to define retention policies, implement legal hold procedures, and maintain export capabilities — ensuring that attestation artifacts are available when needed for legal proceedings, regulatory investigations, or insurance claims.

C.4 Consent Attestation and Healthcare AI

Design Decision: OVERT HITL-1 requires cryptographic attestation of patient consent in healthcare AI deployments, with consent receipts that are independently verifiable.

Rationale: AI systems that generate their own compliance records without actual human attestation create a novel and dangerous category of false documentation.

Supporting Analysis: Recent class-action litigation regarding an ambient clinical documentation system deployed without all-party consent illustrates this risk. The complaint alleged violations of the California Invasion of Privacy Act (CIPA) and the Confidentiality of Medical Information Act (CMIA).

The most significant allegation: the AI tool allegedly inserted false statements into patient charts claiming patients "were advised" and "consented" to recording when they had not. This represents AI systems generating their own false compliance documentation — a pattern that conventional audit methods (reviewing the documentation itself) cannot detect, since the documentation asserts the very compliance whose absence it conceals.

Estimates suggest 100,000+ patient encounters may have been affected. The legal theories — wiretapping, unauthorized third-party disclosure, false consent documentation, retention failures — represent patterns emerging across healthcare AI deployments.

OVERT consent attestation addresses this by requiring that consent events be cryptographically attested with independent verification: the consent receipt is signed by the notary network, not generated by the AI system itself. The receipt proves that a consent interaction occurred at a specific time, was recorded through a specified mechanism, and was attested by an independent party.

This approach also responds to the California Invasion of Privacy Act's requirement for "all party" consent to recording, as well as SB 53's incident reporting requirements effective January 1, 2026.

C.5 Multi-Agent Trust Exploitation

Design Decision: OVERT Sections 11-16 (Agentic AI Controls) require per-call attestation, capability-based access control, and multi-agent trust boundary enforcement.

Rationale: Multi-agent AI systems exhibit systematic vulnerability to trust exploitation through prompt injection, tool misuse, and cross-agent privilege escalation.

Supporting Analysis: Research has demonstrated that multi-agent systems exhibit an 82.4% vulnerability rate to trust exploitation attacks, including: prompt injection through shared context, capability escalation via delegated tool access, and information exfiltration through inter-agent communication channels. The CaMeL framework (Google DeepMind, 2025) proposed capability-based prompt injection defense through separation of privileged and quarantined execution contexts — a design pattern that OVERT formalizes through per-call attestation and capability-scoped access control.

The "policy-quality gap" is particularly acute in multi-agent systems: an attestation system that faithfully records and attests to the outputs of a compromised agent produces cryptographically valid records of invalid outputs. OVERT addresses this through capability-based access control (CAP-1, CAP-2) that constrains what each agent can do, combined with per-call attestation (TOOL-1 through TOOL-5) that records what each agent actually did. The combination enables forensic reconstruction of multi-agent interactions and detection of capability violations even when individual agent outputs are compromised.

C.6 Tiered Certification Analogy

Design Decision: OVERT is structured as a tiered standard (AAL-1 through AAL-4) with progressive requirements, permitting organizations to adopt attestation incrementally while establishing a clear ceiling (AAL-4) for the highest assurance tier.

Rationale: Tiered certification avoids all-or-nothing adoption barriers and creates a progressive path toward comprehensive governance. Building-certification systems provide a useful structural analogue.

Supporting Analysis: The relevant lesson from tiered certification systems is structural, not economic: progressive assurance levels can lower adoption friction, and separation between a standard-setter and an independent verifier can improve trust in published claims. For OVERT, the relevant point is narrower: tiered adoption and separation between standard-setting and independent verification can accelerate uptake without changing the standard's underlying technical claims.

External regulatory, contractual, or market incentives may influence adoption, but they do not alter what OVERT itself proves.

Cautionary lessons: Tiered systems can incentivize point gaming or over-interpretation of lower-tier certifications. For AI governance, this informed OVERT's emphasis on distinguishing documentation, operator-controlled telemetry, and independently verifiable attestation rather than treating all conformance levels as equivalent.

C.7 PCI-DSS Contractual Adoption Precedent

Design Decision: OVERT includes a signal architecture and independent-verification model that can be incorporated into contractual and oversight processes, paralleling the way PCI-DSS was operationalized through private agreements.

Rationale: Standards adoption often accelerates when contractual incentives align with verification requirements.

Supporting Analysis: PCI-DSS illustrates that private agreements can embed verification expectations into commercial relationships without waiting for legislation. The relevant point for OVERT is that procurement, platform, insurance, or sector-specific contracts may reference verifiable governance evidence and independent verification artifacts. OVERT is designed to be referenceable in those settings, but it does not prescribe a particular market structure, assessment industry, or contractual model.

C.8 FedRAMP and NIST SP 800-53 Adoption History

Design Decision: OVERT includes crosswalks to NIST SP 800-53 Rev 5 and FedRAMP (Section 27) and supports OSCAL-formatted attestation packs.

Rationale: Federal adoption of security standards follows established patterns through NIST framework alignment, FedRAMP authorization, and OSCAL-based automation.

Supporting Analysis: Federal security standards often spread through crosswalks, machine-readable artifacts, and reuse in existing compliance workflows. The relevant point for OVERT is interoperability: by mapping to NIST/FedRAMP concepts and supporting OSCAL-compatible outputs, implementers can present attestation evidence in familiar oversight formats. These references explain integration paths, not adoption forecasts or official endorsement.

C.9 Insurance Market Interpretation

Design Decision: OVERT Section 4.6 (Risk Signal Architecture) and Annex D (Risk Signal Framework) are primary design pillars, not afterthoughts.

Rationale: Insurance market reactions illustrate that external risk bearers may seek more verifiable runtime evidence for AI systems. Those reactions are informative context; they do not determine the scope or legal effect of this standard.

Supporting Analysis: Insurance markets have begun issuing both exclusions and affirmative products addressing AI risk. Those developments are relevant here only as evidence that some external risk bearers are beginning to differentiate among AI governance postures; they do not imply insurer endorsement of OVERT, any required coverage position, or any specific underwriting outcome.

These developments show why independently verifiable runtime evidence may become relevant to external risk assessment. OVERT provides a signal and evidence architecture that can be evaluated in that context, but it does not determine coverage availability, pricing, or legal entitlement.

C.10 Non-Egress Architecture and Business Associate Agreement Exposure

> *This section describes architectural properties relevant to regulatory analysis. It does not constitute legal advice. Organizations SHALL obtain qualified legal counsel regarding data processing agreements and BAA requirements for their specific deployments.*

Design Decision: Section 17.5 states that the non-egress architecture "SHOULD be designed to prevent the transmission of Protected Health Information (PHI) or other regulated content" while explicitly hedging: "The applicability of data processing agreements or Business Associate Agreements remains a question of applicable law and regulatory interpretation."

Rationale: The architectural argument for reduced BAA exposure is strong but the legal conclusion is not yet settled. OVERT preserves the argument without overclaiming.

Supporting Analysis: Under HIPAA, a Business Associate is any person or entity that "creates, receives, maintains, or transmits" PHI on behalf of a covered entity (45 CFR §160.103). The OVERT non-egress architecture is specifically designed so that the attestation layer — including the receipt

service, notary network, and transparency log — never receives PHI. Only cryptographic commitments (HMAC-SHA256 digests with tenant-scoped keys) cross the operator's trust boundary. The raw content remains in the operator's content-addressable storage, never leaving the covered entity's environment.

The architectural claim is: if the attestation provider never receives, creates, maintains, or transmits PHI — receiving only irreversible cryptographic commitments from which PHI cannot be reconstructed — the attestation provider may not meet the statutory definition of a Business Associate. This may reduce the factual basis for treating the attestation provider as a recipient of PHI, but BAA obligations remain a matter of applicable law and deployment-specific facts.

However, OCR has not issued guidance specifically addressing whether receipt of cryptographic commitments derived from PHI constitutes "receiving" PHI. The closest analogue is the de-identification safe harbor (45 CFR §164.514(b)), which permits disclosure of health information from which specified identifiers have been removed. HMAC commitments are arguably stronger than de-identification: they are computationally irreversible without the tenant-scoped key, which the attestation provider never possesses.

The hedge in Section 17.5 reflects the current state: the architectural argument is sound, the legal conclusion requires either OCR guidance or judicial interpretation, and the standard should not assert a legal conclusion that applicable law has not yet confirmed. Healthcare operators should consult qualified HIPAA counsel regarding their specific deployment architecture.

C.11 Emergent Behavior in Authorized Agentic Systems

Design Decision: OVERT Section 16 (Behavioral Drift Governance) introduces five control families (DRIFT-1 through DRIFT-5) addressing emergent behavioral changes in agentic AI systems that occur entirely within authorized operational bounds.

Rationale: Existing governance frameworks — including earlier per-action runtime control models — are designed to detect and prevent policy violations: individual actions that breach a defined rule. Agentic AI systems introduce a qualitatively different governance challenge: emergent behavior where every individual control passes but the system's aggregate behavior drifts, cascades, or produces ungovernable complexity. This gap cannot be closed by tightening existing controls; it requires a new category of governance capability.

Supporting Analysis: Per-action attestation — the foundational model used by earlier runtime-governance designs and comparable frameworks — operates on a premise inherited from conven-

tional access control: that governance is the sum of individual authorization decisions. This premise holds for request-response systems where each invocation is independent. It does not hold for agentic AI systems, where persistent agents accumulate state, spawn subordinate agents, and operate across extended time horizons. Six illustrative scenarios demonstrate the structural inadequacy of per-action governance for agentic deployments.

Spawn chain complexity. An orchestrator agent, operating within its declared capability set, spawns sub-agents to decompose a complex task. Each sub-agent, also operating within its declared capability set, spawns further sub-agents. Every individual spawn decision is authorized under the system's capability policy. The resulting execution graph, however, may comprise dozens or hundreds of leaf agents operating in parallel, each issuing tool calls, consuming resources, and producing outputs that feed into sibling and parent agents. The aggregate topology — the total number of active agents, the depth of the spawn hierarchy, the fan-out at each level — may far exceed what any human operator anticipated or any governance process was designed to oversee. Existing controls such as MULTI-2 attest the topology of agent hierarchies but do not evaluate whether the observed topology complexity exceeds the deployment's declared operational baseline.

Within-bounds behavioral drift. An agent produces outputs that individually conform to all applicable policy constraints across successive operational epochs. No single output is flagged, rejected, or escalated. Over time, however, the statistical distribution of those outputs shifts: risk scores trend higher or lower, topic coverage narrows, tool selection patterns change. The shift may be gradual enough that no individual epoch-to-epoch comparison triggers concern, yet the cumulative drift from the system's initial behavioral baseline is substantial. Measurement and evaluation controls such as MEA-2 assess whether individual outputs violate policy thresholds; they do not detect distributional shifts in the population of authorized outputs. Such drift may indicate model degradation, subtle prompt manipulation that biases rather than violates, or environmental changes that alter the agent's effective decision-making.

Cascading depth exploitation. Consider a three-level agent hierarchy where each level spawns three sub-agents. The resulting execution graph contains twenty-seven leaf agents. Each individual agent operates within its authorized bounds — its tool calls are permitted, its outputs conform to policy, its resource consumption falls within declared limits. But the combinatorial complexity of the full execution graph — the total volume of tool invocations, the interaction patterns between agents at different levels, the aggregate resource consumption, the effective attack surface — may be orders of magnitude beyond the deployment's design assumptions. Existing recursion depth limits operate per-trace and do not evaluate the aggregate complexity of concurrent execution graphs sharing a common orchestrator.

Tool selection drift. An agent authorized to invoke multiple tools shifts its selection distribution over time. Where the agent previously selected one tool for approximately sixty percent of invocations

and another for approximately forty percent, the ratio gradually inverts. Neither tool is prohibited; every individual invocation is authorized. The change in selection distribution, however, may indicate that the agent's underlying decision-making behavior has materially changed. Existing controls log individual tool invocations but do not track selection distributions across tools over time, and therefore cannot detect distributional shifts that leave every individual action compliant.

Propagated drift across agent hierarchies. When a parent agent drifts in the manner described above, its outputs — which serve as inputs to downstream agents — change in distribution. Child agents, whose models, policies, and configurations remain unchanged, alter their behavior in response to the changed input distribution. The behavioral drift propagates through the attestation DAG without any agent individually violating its policy. Existing controls evaluate each agent's behavior independently and do not correlate behavioral changes across parent-child attestation linkages, rendering propagated drift invisible to per-agent governance.

Human oversight quality degradation. Human reviewers responsible for overseeing AI outputs initially conduct substantive reviews: they spend adequate time, apply corrections at rates consistent with the system's risk signals, and demonstrate decision patterns that correlate with output characteristics. Over time — through automation bias, workload pressure, or miscalibrated trust — review duration decreases, modification rates decline, and the statistical correlation between risk signals and review decisions weakens. The review process continues to occur, and existing controls attest that it occurred, but the review ceases to be substantively meaningful. Approval velocity controls may cap the rate of approvals but do not assess whether the cognitive engagement underlying each approval is sufficient for the decision's risk level.

These six scenarios share a common structural feature: the governance gap lies not in any individual action but in the relationship between per-action compliance and system-level behavior. An attestation system that evaluates each action independently and finds no violation may nonetheless fail to detect that the system's aggregate behavior has materially changed — potentially in ways that alter its risk profile, undermine its fitness for purpose, or erode the effectiveness of human oversight. DRIFT-1 through DRIFT-5 close this gap by requiring that conformant systems declare their intended behavioral baseline (DRIFT-1), detect deviations from that baseline using sequential statistical methods (DRIFT-2), evaluate execution topology complexity against declared bounds (DRIFT-3), trace behavioral drift propagation across agent hierarchies (DRIFT-4), and assess the substantive quality of human oversight processes (DRIFT-5). The standard specifies what conformant systems must detect and attest. The specific statistical methods, evaluation instruments, and enforcement mechanisms are specified in the registered Protocol Profile.

Annex D: Risk Signal Framework (Informative)

This annex describes the framework for OVERT risk signals. Signal definitions, mathematical formulas, derivation procedures, and minimum credibility thresholds are specified in the registered Protocol Profile or companion signal specification.

D.1 Signal Properties

All OVERT risk signals SHALL satisfy the properties specified in Section 4.6:

1. Content-free derivation
2. Verifiability classification
3. Temporal granularity
4. Statistical rigor
5. Scope binding

D.2 Signal Categories

OVERT risk signals are organized into three categories:

Category	Scope	Examples
Operational Signals	Attestation infrastructure health	Coverage ratios, exposure windows, response latency, retention integrity
Governance Risk Signals	Policy compliance indicators	Violation rate bounds, override frequency, consent coverage, review completion
Agentic Risk Signals	Agentic system behavioral indicators	Behavioral drift rate, graph complexity, spawn authorization, review quality

Agentic Risk Signals apply only to systems classified as "Automation" or "Agentic" under IDE-1.2 and are required for OVERT Agentic conformance.

D.3 Signal Derivation Requirements

Signal specifications in the registered Protocol Profile are required to include, for each signal (see Section 4.6):

- **Signal identifier** — unique, namespaced (e.g., OVERT-INS-NNN for insurance signals, or other prefixes as defined by companion specifications)
- **Definition** — precise natural-language description
- **Formula** — mathematical formula with defined numerator, denominator, and unit
- **Data type and unit** — including encoding requirements to avoid floating-point variance
- **Source artifacts** — which attestation artifacts are required for computation
- **Derivation procedure** — step-by-step auditor-reproducible computation method
- **Aggregation window** — temporal scope (epoch, daily, policy-period)
- **Missing-data handling** — behavior when source artifacts are unavailable
- **Minimum credibility threshold** — minimum sample size for statistically credible interpretation
- **Severity classification** — threshold-based severity levels

D.4 Design Rationale

Risk signals are a primary design goal of OVERT because the verification gap described in the Foreword affects defenders, auditors, regulators, procurement reviewers, and external risk assessors alike. Quantitative risk signals — independently verifiable where the denominator source supports it, operator-dependent where it does not (see Section 4.6) — enable:

- **Security operations:** Monitoring of coverage, overrides, exposure windows, and other runtime indicators within the attested scope
- **Audit and investigation:** Recomputable evidence for control-execution and anomaly analysis
- **Regulatory and oversight reporting:** Quantified posture reporting without content exposure
- **External risk analysis:** Structured inputs for insurance, procurement, or other third-party evaluations, subject to the verifiability classification of the underlying signals

Signal specifications are maintained in the Protocol Profile rather than this standard so that signal definitions can evolve with operational experience and measurement practice, without requiring standard revisions.

Annex E: Legal Admissibility Analysis (Informative)

> *This annex is informative only and does not constitute legal advice. Admissibility determinations are made by courts applying jurisdiction-specific rules. Organizations should consult qualified legal counsel regarding the admissibility of attestation artifacts in their jurisdictions.*

This annex analyzes how AAL-4 attestation artifacts produced by OVERT-conformant systems may relate to evidentiary rules governing the admissibility of electronic records. The discussion identifies how OVERT design features address foundational admissibility concepts — authenticity, integrity, chain of custody, and hearsay exceptions — without asserting that any specific attestation artifact will be admitted in any specific proceeding.

E.1 Federal Rules of Evidence 902(13): Certified Records of Regularly Conducted Activity (Electronic)

Rule: FRE 902(13), effective December 1, 2017, provides for self-authentication of "a record of a regularly conducted activity" in electronic form, when accompanied by a certification from a qualified person that the record: (A) was made at or near the time of the occurrence of the matters set forth by a person with knowledge, or from information transmitted by such a person; (B) was kept in the course of the regularly conducted activity; and (C) was made as a regular practice of that activity.

OVERT Mapping:

FRE 902(13) Requirement	OVERT Feature	Relevant Controls
"made at or near the time of the occurrence"	Attestation receipts include nanosecond-precision timestamps (wall_time_ns), co-epoch binding, and transparency log inclusion with signed tree heads providing independent temporal verification.	ATT-1, ATT-2, Section 18

FRE 902(13) Requirement	OVERT Feature	Relevant Controls
"by a person with knowledge, or from information transmitted by such a person"	OVERT records are machine-generated by the arbiter and notary network. Courts increasingly accept automated systems as sources of business records when the system's reliability is established. The notary network's independent derivation of binary identity and network state provides an additional reliability indicator.	ATT-2.2, ATT-3.3, ATT-5
"kept in the course of regularly conducted activity"	OVERT attestation is continuous and automatic — operating on every in-scope AI action as a regular practice, not created in anticipation of litigation. The transparency log provides a tamper-evident, append-only record.	ATT-4, Section 21.1
"made as a regular practice"	AAL-4 conformance requires continuous attestation for all in-scope actions. The DPL publishes request commitments per epoch as a regular operational practice.	Section 22 (Conformance)
Certification by qualified person	Section 21.3(e) (Custodian Certification) requires the export package to include custodian identity, timestamp, scope declaration, and hash of the export package. This certification can be prepared by the operator's designated custodian of records.	Section 21.3, 21.5

Analysis: OVERT attestation artifacts are designed to address the structural elements of FRE 902(13). Whether a specific court accepts these artifacts under FRE 902(13) will depend on the proponent's compliance with notice requirements (FRE 902(13) requires written notice and opportunity to inspect), the court's assessment of the underlying system's reliability, the proponent's ability to establish the "regular practice" and "person with knowledge" elements for machine-generated records, and the specific facts of the deployment. This analysis identifies design alignment; it does not predict admissibility outcomes.

E.2 Federal Rules of Evidence 902(14): Certified Data Copied from Electronic Device, Storage Medium, or File

Rule: FRE 902(14), also effective December 1, 2017, provides for self-authentication of data "copied from an electronic device, storage medium, or file" when accompanied by a certification from a qualified person that the process of digital identification used to verify the data is trustworthy, typically through cryptographic hash verification.

OVERT Mapping:

FRE 902(14) Requirement	OVERT Feature	Relevant Controls
Data "copied from" electronic device	OVERT immutable export packages (Section 21.3) are copied from the operator's content-addressable storage and the transparency log.	Section 21.3
Process of digital identification	SHA-256 hashing, HMAC commitments, Ed25519/BLS signatures, and Merkle tree inclusion proofs provide multiple layers of cryptographic verification.	ATT-1, Section 18, Annex B
Process "used to verify" is trustworthy	The entire OVERT verification chain is publicly specified, uses NIST-approved cryptographic primitives (SHA-256, HMAC-SHA256, HKDF), and is independently reproducible by any party.	Protocol Profile, Annex B
Certification by qualified person	Section 21.3(e) requires custodian certification with identity, timestamp, scope, and hash of the export package.	Section 21.3(e)

Analysis: FRE 902(14) was specifically designed to accommodate cryptographic hash verification of electronic data. OVERT attestation artifacts employ multiple layers of cryptographic integrity verification (content hashes, commitment chains, Merkle tree proofs, notary signatures) designed to address the requirements of 902(14) authentication. The publicly documented verification procedures enable opposing parties and courts to assess the trustworthiness of the digital identification process.

E.3 Federal Rules of Evidence 803(6): Business Records Exception to Hearsay

Rule: FRE 803(6) excludes from the hearsay rule a record of a regularly conducted activity if: (A) made at or near the time by someone with knowledge; (B) kept in the course of a regularly conducted business activity; (C) making the record was a regular practice; and (D) these conditions are shown by testimony of the custodian or another qualified witness, or by a certification under FRE 902(11), (12), or (13). The record may be excluded if "the source of information or the method or circumstances of preparation indicate a lack of trustworthiness."

OVERT Mapping:

Elements (A), (B), (C), and (D) map identically to the FRE 902(13) analysis above. The additional trustworthiness inquiry under FRE 803(6)(E) is addressed by OVERT's design properties:

Trustworthiness Factor	OVERT Feature
Source reliability	Attestation records are generated by cryptographically verified arbiters (binary identity derived by independent notaries, not self-reported) operating within verified network isolation (NETATT).
Preparation circumstances	Attestation is continuous, automatic, and not created in anticipation of litigation. The system operates identically regardless of whether litigation is pending.
Tamper evidence	Transparency log provides append-only storage with Merkle tree consistency proofs. Any modification is detectable through signed tree head comparison.
Independent verification	Any party can independently verify attestation integrity using publicly available verification procedures without operator cooperation.

Analysis: OVERT attestation artifacts are designed with properties relevant to the FRE 803(6) trustworthiness inquiry. Whether a specific court finds a particular deployment's attestation artifacts trustworthy under 803(6)(E) will depend on the deployment-specific facts, including system reliability, operational consistency, and the circumstances of record creation. This analysis identifies design alignment; it does not predict admissibility or trustworthiness determinations.

E.4 Federal Rules of Civil Procedure 37(e): Failure to Preserve ESI

Rule: FRCP 37(e) addresses the consequences of failing to preserve electronically stored information (ESI) that should have been preserved in anticipation or conduct of litigation. If ESI is lost because a party failed to take reasonable steps to preserve it, and it cannot be restored or replaced through additional discovery, the court may: (1) upon finding prejudice, order measures no greater than

necessary to cure the prejudice; or (2) upon finding that the party acted with intent to deprive, presume the information was unfavorable, instruct the jury accordingly, or dismiss the action.

OVERT Mitigation:

OVERT Section 21 (Legal Preservation and Production) directly addresses FRCP 37(e) exposure:

FRCP 37(e) Element	OVERT Mitigation	Relevant Section
"reasonable steps to preserve"	Section 21.1 requires operators to define and publish retention policies meeting or exceeding the longer of regulatory requirements or applicable statutes of limitation. Section 21.2 requires legal hold procedures upon receipt of litigation hold notice or preservation demand.	Section 21.1, 21.2
"lost because a party failed"	The transparency log provides an independent, tamper-evident record of what attestation artifacts existed. Even if the operator's local copy is lost, the transparency log entries (hashes, inclusion proofs) remain, proving that the artifacts existed and establishing their content hashes.	ATT-4, ATT-4 (Section 8)
"cannot be restored or replaced"	Section 21.3 requires immutable export capabilities. The transparency log + notary signatures provide partial reconstruction capability even if local evidence is lost. Co-epoch binding and receipt hashes enable a court to determine the scope of loss.	Section 21.3, 21.4
"intent to deprive"	OVERT audit trails make intentional destruction detectable. If an operator deletes local evidence, the transparency log still contains the receipts — showing what was attested and when. The gap between transparency log entries and available local evidence is itself evidence of deletion.	ATT-4, ATT-4 (Section 8)

Analysis: OVERT does not eliminate FRCP 37(e) exposure — no technical system can prevent a party from destroying evidence if they are willing to accept the consequences. However, OVERT creates a structural environment where: (a) preservation obligations are documented in the operator's published retention policy; (b) legal hold procedures are defined and attestable; (c) the transparency log provides an independent record of what artifacts existed, making destruction detectable; and (d) the gap between what the log shows existed and what the operator can produce is itself a measurable, verifiable retention-integrity signal.

E.5 International Admissibility References

United Kingdom: Civil Evidence Act 1995

The Civil Evidence Act 1995 abolished the common law rule against hearsay in civil proceedings, making all relevant evidence admissible subject to weight. Section 8 provides that where a statement in a document is produced by a computer, the statement may be proved by "a certificate identifying the document and signed by a person occupying a responsible position in relation to the operation of the computer." OVERT attestation artifacts, accompanied by custodian certification (Section 21.3(e)), are designed to support certification under Section 8. The weight given to such evidence remains at the court's discretion, informed by factors including the reliability of the computer system and the manner in which the data was processed.

The UK has not enacted standalone AI legislation as of March 2026, maintaining a "pro-innovation" regulatory approach with cross-sector principles applied through existing regulators. The Online Safety Act 2023, with a February 2026 amendment bringing standalone AI chatbots within scope, creates an expanding statutory backdrop. In the absence of AI-specific evidentiary rules, OVERT attestation artifacts would be assessed under general principles of electronic evidence admissibility.

European Union: eIDAS Regulation (Regulation 910/2014 and eIDAS 2.0)

The eIDAS Regulation provides a legal framework for electronic identification and trust services across EU member states. Under eIDAS:

- **Electronic signatures** (Article 25): An electronic signature shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic signatures have the equivalent legal effect of handwritten signatures.
- **Electronic seals** (Article 35): An electronic seal shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic seals enjoy a presumption of integrity of the data and correctness of the origin.
- **Electronic time stamps** (Article 41): An electronic time stamp shall not be denied legal effect solely on the grounds that it is in electronic form. Qualified electronic time stamps enjoy a presumption of accuracy.
- **Electronic documents** (Article 46): An electronic document shall not be denied legal effect solely on the grounds that it is in electronic form.

OVERT attestation artifacts — cryptographically signed, timestamped, and integrity-verified — are designed to satisfy the structural requirements for electronic evidence under eIDAS. Organizations operating under eIDAS may additionally seek qualified trust service provider (QTSP) status for their

notary network operations, which would provide the legal presumptions associated with qualified electronic signatures, seals, and time stamps. The eIDAS 2.0 update extends the framework to include electronic ledgers, which may be relevant to OVERT transparency log operations.

The revised EU Product Liability Directive (Directive 2024/2853), with transposition deadline December 9, 2026, explicitly treats software and AI systems as products. Article 9 creates rebuttable presumptions of defectiveness where a defendant fails to comply with disclosure obligations. OVERT attestation artifacts are designed to support the operator's ability to respond to disclosure obligations with verifiable, tamper-evident records.

Annex F: Sample Citation Language (Informative)

This annex provides canonical citation forms for referencing OVERT conformance in legal, procurement, insurance, and regulatory contexts.

F.1 Canonical Conformance Citation Format

The standard citation form for an OVERT conformance claim follows the grammar defined in Section 22.4. All claims include a human-readable scope summary and exclusions summary. Level 3 and Level 4 claims additionally include coverage percentage, denominator source and verifiability classification, scope hash, and exposure-window duration.

Example (Level 3):

> *OVERT Level 3 Agentic — v1.0.0, Protocol Profile 1.0, Scope Summary: sys-agent-010 patient-facing agentic workflows (API gateway gw-prod-01, FHIR interface, voice endpoint), Exclusions: None (full coverage verified), Scope: 85% of inbound API traffic, Denominator: Independent, Scope Statement: sha256:[scope-hash], Exposure Window: 0h (0%), IAP Topology: Multi-IAP, as of 2026-06-15*

Example (Level 2):

> *OVERT Level 2 Core — v1.0.0, Profile v1.0, Scope Summary: sys-cda-001 clinical documentation API (FHIR R4 interface, HL7v2 ADT feed), Exclusions: Not assessed: batch-analytics-002 (scheduled for Q3 assessment), as of 2026-03-15*

F.2 Guidance for Referencing OVERT in External Documents

Organizations referencing OVERT conformance in procurement, insurance, regulatory, or legal contexts SHOULD:

1. Use the canonical citation format from Section F.1, which includes scope summary, exclusions, denominator source, and exposure-window fields.
2. Not imply that OVERT conformance establishes legal compliance, regulatory approval, insurance coverage, or a judicially recognized standard of care.
3. Not imply that OVERT conformance covers systems, interfaces, or traffic classes outside the declared scope.
4. Consult qualified legal counsel when drafting contract, insurance, or regulatory language that references OVERT.
5. Clearly distinguish between independently verifiable signals and operator-dependent signals when making claims about evidence quality.

> ***NOTE*** *— Previous versions of this annex included sample legal, procurement, insurance, and regulatory paragraphs. Those samples were removed because copy-paste-ready advocacy language in a standard creates misrepresentation risk. Organizations should draft context-specific language with qualified counsel using the canonical citation format and the scope/exclusions/denominator disclosures required by Section 22.4.*

F.3 Disclaimer

> ***NOTE*** *— This annex is informative only and does not constitute legal advice. OVERT has not been judicially recognized as defining a standard of care. No insurer, regulator, or court has adopted OVERT as dispositive evidence. Citation forms should be adapted to the specific legal jurisdiction, regulatory framework, and contractual context in which they are used. Organizations should consult qualified legal counsel when referencing OVERT in any external document.*