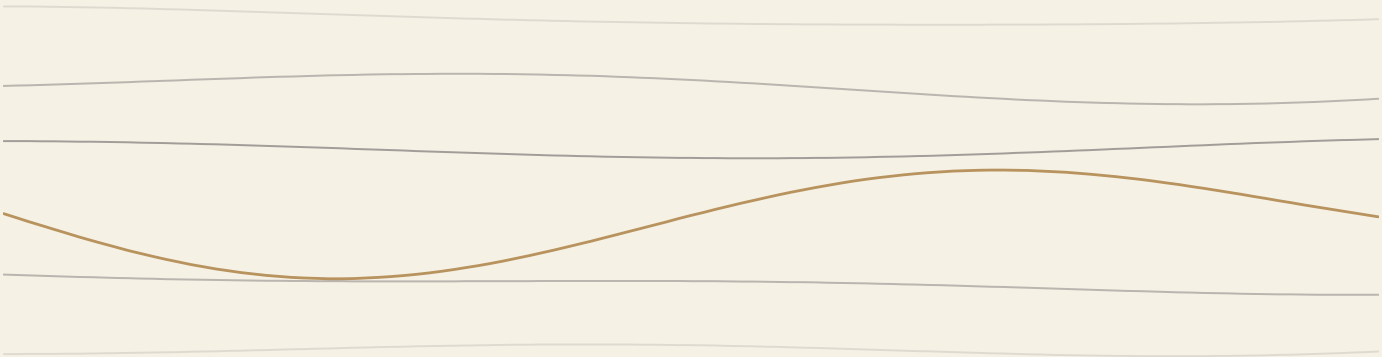
OVERT

OBSERVABLE VERIFICATION EVIDENCE FOR RUNTIME TRUST



What Good Governance Must Prove

Six domains — Govern · Identify · Protect · Attest · Measure · Respond

DATE	June 2026
PUBLISHED BY	GLACIS Technologies, Inc.
REPRODUCES	Part 2: Governance Domains (Sections 5–10)
COMPLETE EDITION	overt.is
CONTACT	overt-review@glacis.io

OFFPRINT NOTICE

This fascicle reproduces Part 2: Governance Domains (Sections 5–10) of OVERT Version 1.1 without modification. Section numbering follows the Complete Edition, which is the sole authoritative text for conformance purposes. Conformance claims cite OVERT 1.1, never an individual fascicle. The Complete Edition and all fascicles are published at overt.is.

Contents of this Volume

PART 2: GOVERNANCE DOMAINS

> ***Six obligations turn an intention to govern into evidence that it happened — govern, identify, protect, attest, measure, respond. A domain is not satisfied by having a control; it is satisfied by being able to prove the control ran.***

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.

Where a control table in Part 2 or Part 3 restates an architecture requirement specified in Part 4, the Part 4 specification prevails in any conflict. Control tables summarize; Part 4 defines.

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
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

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

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

ID	Control	Attestation Artifact	Level
GOV-3.2	Assign severity levels to each risk category with escalation criteria	Severity matrix with escalation procedures	AAL-2
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., the NIST AI RMF Generative AI Profile (NIST AI 600-1), 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

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: notifica-	Contract excerpts or attestation from legal review	AAL-2

ID	Control	Attestation Artifact	Level
	tion of material changes, incident disclosure, and cooperation with audits		

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) + AAL-4 (receipt reference in response metadata, GOV-5.6)

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. Where AI-generated content carries C2PA Content Credentials (GOV-5.3), the receipt reference SHOULD be embedded as a C2PA assertion, so that any downstream consumer of the content can walk from the artifact to its attested enforcement record — end-to-end output provenance	Receipt reference in response metadata; dispute resolution procedure; C2PA assertion where content is labeled	AAL-4

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

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

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

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, at minimum: the effective egress policy, the enforcement component identity, and the TLS certificate pin set (Section	Co-epoch NETATT covering the Section 18.3 minimum set	AAL-4

ID	Control	Attestation Artifact	Level
	18.3). Operators MAY include additional deployment-specific inputs — network policy definitions (hashing the policy controller input, not dynamic ephemeral rules), network controller identity, eBPF programs, CNI configuration, runtime environment variables affecting AI behavior. The minimum input set is specified in the registered Protocol Profile; Section 18.3 prevails in any conflict		
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

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

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

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; closed schema (unknown fields rejected)	AAL-4

ID	Control	Attestation Artifact	Level
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
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 insuffi-	Notary-derived binary identity in receipt	AAL-4

ID	Control	Attestation Artifact	Level
	cient for AAL-4 conformance. See Section 18.2 for acceptable measurement pipelines		
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
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, or pure 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

ID	Control	Attestation Artifact	Level
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
ATT-3.5	**Optimistic Enforcement Mode:** Because Phase 3 notary counter-signature is asynchronous by architecture (ATT-3.3), execution proceeds after Phase 2 (Provisional Receipt). The normative provisions (a)–(f) following this table govern which actions may proceed on a provisional receipt alone: high-risk action classes SHALL carry independent authorization before execution, and the optimistic-residue caps bound what may execute with neither	Optimistic mode declaration in policy with explicit tool-call classification; capability artifacts or synchronous attestation for high-risk classes; circuit breaker on notary rejection	AAL-4

ATT-3.5 normative provisions. Phase 3 notary counter-signature is asynchronous (ATT-3.3); execution after Phase 2 is therefore the universal execution model of this standard. The provisions below govern what may proceed on a provisional receipt alone, and bound the **optimistic residue** — actions that execute with neither independent pre-authorization nor synchronous attestation.

(a) **Action classification and independent authorization.** Optimistic execution on Phase 2 alone is permitted for actions classified Read-only, subject to output containment: a Read-class action that returns sensitive content (PII, PHI, classified data, or credentials) remains optimistic-eligible only where delivery is to the authenticated, authorized subject of the session over a sink declared in the mediation scope statement. Routing of sensitive content to any undeclared or uncontrolled sink SHALL be blocked regardless of action classification — the restriction binds on exfiltration paths, not on governed delivery to the authorized subject. Actions classified Write, Transact, Delete, or Modify SHALL carry independent authorization before execution, satisfied by either: (i) a valid capability artifact (3.24) covering the tool or action class, the session, and the time of execution, verified at the enforcement point without a network dependency; or (ii) synchronous Phase 3 notary attestation, where the deployment elects to block. An action in these classes executed with neither constitutes optimistic residue and is subject to provisions (c) through (f). Misclassification of a side-effecting action as Read-only is a governance failure and SHALL be reported as a conformance deviation.

(b) **Circuit breaking.** If Phase 3 subsequently rejects a provisional receipt (notary detects drift, binary mismatch, or policy violation), the system SHALL trigger a circuit breaker (TOOL-3.3) halting subsequent requests from the same agent or 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 executed as optimistic residue during the claimed period SHALL be disclosed in the conformance statement, as both the claim-period average and the worst single-epoch value.

(e) **Residue caps.** The caps below apply to the optimistic residue, independently per action class (Read, Write, Transact, Delete, Modify, and any operator-defined classes). Actions executed under a valid capability artifact or under synchronous Phase 3 attestation are independently authorized and do not count toward the residue.

Conformance Level	Worst Single Epoch	Claim-Period Average
Level 4	25% of in-scope actions	15% of in-scope actions
Level 3	40% of in-scope actions	—

Deployments exceeding a cap for any action class SHALL claim Level 3 for the affected scope until the residue is reduced; a Level 3 deployment exceeding its epoch cap SHALL report the excess as a conformance deviation. This structure ensures that Evidence-Grade conformance reflects predominantly independently authorized or independently verified execution, and that optimistic execution cannot be concentrated in high-risk epochs and diluted by low-risk traffic.

(f) **Eligibility governance.** The operator's action classifications and capability-issuance policy SHALL be documented in the mediation scope statement and SHALL be subject to review by the IAP or auditor upon request.

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
ATT-4.3	Provide consistency proofs between any two Signed Tree Heads	Merkle consistency proof	AAL-4

ID	Control	Attestation Artifact	Level
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 standard. Publish governance model to transparency log	Governance model documentation with transparency log proof	AAL-4

ID	Control	Attestation Artifact	Level
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

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

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

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.]