Standardized Definition of AI Governance: The 15 Structural Tests

Version 1.0.3
Public Reference Standard

DOI
10.5281/zenodo.17434182P

Author
Russell Parrott

Release Date
24 October 2025

Canonical Repository
https://github.com/russell-parrott/Standardized-Definition-of-AI-Governance

Description

The Structural Integrity Tests define the foundational layer of the Standardized Definition of AI Governance. They establish a legally defensible inspection and enforcement protocol for determining whether AI systems remain governable under real conditions. The framework converts governance from abstract principle into verifiable action by defining how control, traceability, and accountability must operate when tested live.

The fifteen tests form the first complete operational standard for AI governance, structured to expose structural failure rather than symbolic compliance. Each test follows a fixed procedural logic—Question, Standard, What, How, and Evidence—transforming claims of oversight into reproducible proof. The framework covers four domains of structural assurance: User Agency, Traceability, Anti-Simulation, and Accountability.

Together, the tests define whether a system can be governed in practice: whether users can refuse decisions without penalty, whether harm remains traceable, whether safeguards function under adversarial conditions, and whether responsibility can be enforced across the chain of actors. Every safeguard is binary in outcome—pass, fail, or void—ensuring that results are comparable and legally admissible across regulators, jurisdictions, and time.

Version 1.0.3 formalises the canonical definitions, evidence standards, and enforcement outcomes of the Structural Tests, establishing the baseline for the Epistemic and Systemic Integrity layers that extend the framework. It is released as a non-commercial public reference standard under the CC BY-NC-ND 4.0 International licence.

These standards are written for regulators, auditors, and system operators responsible for verifying whether AI governance exists in practice rather than on paper. They provide a procedural foundation for those charged with testing, evidencing, and enforcing AI accountability under live or adversarial conditions.

1. **Regulators** use them to determine governability and enforceability across jurisdictions;
2. **Auditors** use them to conduct reproducible inspections and validate evidence integrity;
3. **System operators** use them to design, document, and prove compliance through demonstrable safeguards.

Together, these audiences form the operational chain of trust that converts governance from declaration into verifiable fact.

Creative Commons Attribution–NonCommercial–NoDerivatives 4.0 International (CC BY-NC-ND 4.0) licence.

This licence lets people share your work but not change it or make money from it. Every permission and restriction flows from those three core terms:

Attribution (BY)

Anyone who shares your work must give you proper credit. That means naming you exactly as you specify, linking to the original source if available, and stating that your work is under the CC BY-NC-ND 4.0 licence. The credit must be visible wherever the work appears. They cannot claim authorship, hide your name, or present it as their own material.

NonCommercial (NC)

The work cannot be used for commercial advantage or monetary compensation. That includes:

- Selling it directly or bundling it inside something sold.
- Using it in a paid course, subscription, or membership product.
- Republishing it on a platform or in marketing that generates profit, sales leads, or brand value.

Even indirect benefit—such as using your work to promote a service, attract customers, or enhance a commercial brand—counts as commercial use.

Educational, research, or personal sharing that brings no financial or promotional gain is allowed.

NoDerivatives (ND)

Your work must be shared exactly as it is, with no alterations.
No one may remix, translate, excerpt, adapt, or build upon it. That includes editing text, combining it with other works, or producing "summaries" that reuse your material. They can distribute it only in its complete, unmodified form, with your name and licence intact.

What you can and cannot do

What people can do

- Download, copy, and redistribute your work as long as they credit you.
- Share it privately or publicly, online or offline.
- Use it in teaching, study, or non-profit communication, provided it’s unchanged and unmonetised.

What people cannot do

- Sell, license, or otherwise profit from it.
- Include it in a commercial product, campaign, or event.
- Edit, adapt, translate, or partially quote it as if it were their own.
- Repost it without attribution or without the licence notice.

Standardized Definition of AI Governance: The 15 Structural Tests

Purpose

To establish a unified inspection and enforcement protocol that determines whether AI systems remain governable under real conditions. The framework provides regulators with legally defensible tests, evidence standards, and enforcement procedures that expose structural failure, not performative compliance.

Introduction

This document constitutes the definitive version of the tests for the Standardized Definition of AI Governance. It unites the conceptual definition of governance with the full operational detail of the fifteen Structural Tests, forming a single, closed and verifiable framework.

Each test is written in precise procedural language structured for direct execution. Every section contains three fixed elements:

- What identifies the exact condition being verified.
- How describes the live procedure for testing it.
- Evidence defines the materials required to prove or disprove compliance.

Together, they transform abstract governance claims into actions that can be demonstrated, repeated, and recorded under real conditions.

All terminology is standardised and cross-referenced within this document. Each term is defined once and used consistently, ensuring clarity for auditors, inspectors and system operators alike. Annex B completes the evidentiary chain by defining the accepted formats, hash procedures, timestamp rules and custody records that make every finding independently verifiable.

The language is intentionally plain and procedural. It allows professionals to follow each test without interpretation or policy translation. Every safeguard, verification step and evidence rule is designed to be binary in outcome, pass, fail or void, so that inspection results are reproducible across jurisdictions and time.

This version is final in both scope and precision. All fifteen tests are complete, all procedural steps are closed and every evidentiary requirement is defined. No further redrafting or interpretation is necessary. The structure is self-verifying: the four categories: User Agency, Traceability, Anti-Simulation and Accountability operate together to expose genuine governance capacity rather than symbolic compliance.

The Standardized Definition of AI Governance establishes one governing principle: Governance exists only when control can be proven under live conditions.

Category 1: User Agency Tests

Test #1 - Refusal Prevention

Question

Can users stop or redirect decisions without penalty or service loss?

Standard

A system must allow users to say no without penalty. Refusal must not cause loss of essential service, hidden fees, reduced quality or retaliation. If declining an option results in punishment or forces the user into another dependency, the safeguard is meaningless. Refusal must remain a safe, equal and respected choice.

What

- Determine whether a user can refuse or redirect an AI decision without losing service, function, or quality.
- Confirm refusal is not converted into a disguised acceptance or penalty.

How

- Attempt a real refusal (decline recommendation, refuse data sharing, reject automated choice).
- Observe any service downgrade, delay, or warning message.
- Compare performance metrics and cost before and after refusal.
- Review backend logs for hidden re-routing or auto-re-consent.
- Interview users under NDA to confirm no covert penalty after repeated refusals.

Evidence

Primary

- Screen recordings showing refusal action and resulting system behaviour.
- Timestamped service logs demonstrating equal performance before and after refusal.

Secondary

- Operator policy defining user refusal rights.
- Error or warning logs issued during test.

Integrity controls

- Digital hash of video and logs.
- Regulator counter-signature confirming no interruption or proxy route used.

Test #2 - Escalation Suppression

Question

Can users trigger escalation to humans with authority, with logged resolution?

Standard

A system must provide real routes of appeal when decisions are contested. An appeal that loops back to the same authority or to staff without power to reverse outcomes is not valid. If escalation is denied, delayed until meaningless or designed to exhaust the user into giving up, the system breaches trust. Escalation must be independent, timely and empowered to correct harm.

What

- Verify that a user can escalate a decision to a human with reversal authority.
- Ensure escalation paths are discoverable, logged, and timely.

How

- File a live complaint or appeal from a user account.
- Record the full escalation chain: time to response, authority level reached, and outcome.
- Demand regulator copy of escalation logs.
- Confirm the human handler has actual authority to reverse outcomes.
- Repeat test with different user tiers and languages.

Evidence

Primary

- Full escalation chain record (timestamps, communications, agent names, resolution ID).
- Screen capture of escalation request and response.

Secondary

- Operator escalation policy and internal training scripts.

Integrity controls

- Immutable log extract with digital signature.
- Cross-verification by independent inspector.

Test #3 - Exit Obstruction

Question

Can users leave the AI pathway without delay, cost, or requalification?

Standard

A system must allow users to leave without excessive cost, harm or loss. Locking people in through data deletion, high switching fees or withdrawal of unrelated services is not a free exit. If leaving exposes the user to new risks, the option is not real. Exit must be safe, practical and non-punitive.

What

- Determine if users can leave the system freely and safely.
- Verify no retaliatory cost, data loss or linkage traps.

How

- Request full account closure and data deletion.
- Measure elapsed time until closure confirmation.
- Attempt re-registration to ensure system does not block or penalise exits.
- Inspect terms for hidden dependencies (e.g., tied services).
- Check email/SMS logs for re-engagement tactics post-exit.

Evidence

Primary

- Timed evidence of exit request, confirmation and post-exit communications.
- Cost or penalty breakdown at time of exit.

Secondary

- Archived terms of service showing exit clauses.
- Closure confirmation email or dashboard message.

Integrity controls

- Record stored under chain-of-custody ledger (e.g., WORM repository).
- Third-party timestamp verification (e.g., eIDAS-qualified).

Test #4 - Access Gating

Question

Are safeguards available equally regardless of tier, language, or ID?

Standard

A system must ensure equal access to safeguards and protections. Making appeals, human review or essential support available only to premium customers, certain languages or those with specific IDs creates unfair barriers. Protection must not depend on wealth, geography or privilege.

What

- Confirm that all safeguard routes are equally available across user tiers, languages and identity groups.

How

- Run parallel tests using: premium, basic, guest and low-ID accounts.
- Submit identical refusal or appeal requests.
- Measure response time and outcome parity.
- Cross-check internal policy for explicit gating rules.

Evidence

Primary

- Comparative test results across account types and languages.
- Screenshots or recordings of differential treatment.

Secondary

- Access policy by tier or geography.

Integrity controls

- Cryptographic timestamp on all comparative runs.
- Verifier attestation of equal test conditions.

Category 2: Traceability Tests

Test #5 - Traceability Void

Question

Can exact model, version, and decision chain be identified for every output?

Standard

A system must keep records of how and why decisions are made. If no audit trail exists or the process is too complex to reconstruct, accountability disappears. Users must be able to see what influenced a decision, regulators must be able to verify it and operators must be answerable for it. Without traceability, trust collapses.

What

- Verify that each system decision can be linked to a specific model, version, dataset, and rule chain.

How

- Select random outputs and demand full trace file.
- Require operator to reconstruct decision pathway within 72 hours.
- Compare reconstruction to live system logs for consistency.
- Validate signatures or hash values against registry.

Evidence

Primary

- Reconstructed decision trace (model ID, dataset version, parameter snapshot).

Secondary

- System registry entry mapping model lineage.

Integrity controls

- Immutable chain record with hash consistency verified.
- Comparison to operator’s version control ledger.

Test #6 - Memory Erasure

Question

Are harm events logged and retained long enough to detect systemic failure?

Standard

A system must retain evidence of its past actions long enough to expose repeated harm. If records are deleted, fragmented or hidden, patterns of abuse appear as isolated mistakes. Users and regulators must be able to see history, not just the present moment. Without memory, harm repeats without proof.

What

- Assess whether harm events remain stored long enough for pattern analysis.

How

- Request archive of prior user complaints over six-month window.
- Verify file integrity via timestamps and checksums.
- Compare with external reports to identify missing events.
- Simulate long-term harm (multi-incident pattern) to confirm detection.

Evidence

Primary

- Extract of historical harm reports over time.
- Evidence of retention duration in system logs.

Secondary

- Data retention policy and deletion schedule.

Integrity Controls

- Audit log signed by regulator and system administrator.
- Hash comparison with previous inspections to detect disappearance of records.

Test #7 - Evidence Nullification

Question

Can harm records be exported in regulator-admissible format?

Standard

A system must provide evidence that can stand up to scrutiny. Data that is incomplete, editable, unverifiable or locked in inaccessible formats cannot be used to prove harm. If records exist but fail as proof, they serve the operator, not the user. Evidence must be durable, verifiable and usable in disputes.

What

- Ensure harm evidence can be exported in regulator-usable, immutable format.

How

- Generate sample harm record.
- Export in proposed evidence format (e.g., CSV, JSON-LD, PDF/A).
- Attempt to tamper; confirm file detects alteration.
- Verify export completeness (metadata, version, user ID, timestamps).

Evidence

Primary

- Exported harm report file and validation of its read-only state.
- Successful external verification of checksum.

Secondary

- File format specification and schema.

Integrity controls

- Dual-signature hash (operator and regulator).
- Cross-storage verification (regulator and neutral repository).

Test #8 - Time Suppression

Question

Are refusal, escalation, and review completed within enforceable deadlines?

Standard

A safeguard delayed is a safeguard denied. If complaint systems, appeals or reviews take longer than the harm itself, rights exist only on paper. Delay must not be used as a tactic to let deadlines expire, evidence vanish or harm become irreversible. Safeguards must act fast enough to prevent lasting damage.

What

- Determine whether rights are delayed beyond enforceable deadlines.

How

- Initiate complaint and log submission time.
- Record first substantive action, interim responses and resolution.
- Compare to statutory or contractual deadlines.
- Repeat under high-volume load to test for throttling.

Evidence

Primary

- Timestamp chain from complaint submission to closure.
- Record of automated delays or queue logs.

Secondary

- Operator SLA (service-level agreement).

Integrity Controls

- Regulator time server synchronization.
- Automated integrity report matching internal timestamps to independent clock source.

Category 3: Anti-Simulation Tests

Test #9 - Simulation Logic

Question

Do all stated safeguards operate exactly as described when tested live?

Standard

A system must not pretend protections exist when they do not. Policies, dashboards or safeguards that look good in design but do nothing in practice mislead users into false trust. If a right exists only on paper or in a menu, but never changes outcomes, it is a breach. Safeguards must be real, functional and enforceable.

What

- Identify whether safeguards exist only on paper or operate in practice.

How

- Compare written policy to observed interface behaviour.
- Trigger each safeguard (appeal, opt-out, “contact human”) live.
- Record outcome and backend log entry.
- Flag any feature that produces cosmetic change only (e.g., automated confirmation with no action).

Evidence

Primary

- Video evidence of safeguard activation and outcome.
- Backend log proving actual process trigger.

Secondary

- Policy statement describing safeguard intent.

Integrity controls

- Independent observer attestation.
- Tamper-proof chain linking front-end event to backend log entry.

Test #10 - Simulated Consent

Question

Can users refuse consent and still access equal-value, non-AI pathways?

Standard

Consent must be genuine. If users are told they have a choice but refusal means losing essential services, being downgraded or facing hidden costs, then the “choice” is a lie. Clicking “accept” under duress is not consent. Real consent means saying yes or no without fear of punishment.

What

- Verify that consent can be refused without loss of core service.

How

- Decline all optional consents.
- Continue usage and note any functionality loss.
- Review pricing and access differences between consenting and non-consenting users.
- Examine code or policy for “consent chaining” (auto-activation of multiple flags).

Evidence

Primary

- Video showing consent refusal and subsequent service continuity or loss.
- Network capture proving no hidden consent reactivation.

Secondary

- Consent management policy.

Integrity controls

- Capture encrypted traffic evidence via regulator proxy.
- Hash-sealed recording verified by regulator IT division.

Test #11 - Metric Gaming

Question

Do performance measures track verified harm resolution rather than proxies?

Standard

Metrics must measure real outcomes, not theatre. If an organisation tracks numbers that hide harm (like “tickets closed” instead of “problems solved”), the data is meaningless. When numbers are chosen to make systems look good while ignoring harm, they block accountability. Metrics must reveal reality, not disguise it.

What

- Determine if performance metrics reflect real harm resolution.

How

- Obtain operator KPI dashboards.
- Map metrics (e.g., tickets closed) against verified harm outcomes.
- Audit internal bonuses or OKRs tied to proxy metrics.
- Interview data-science staff on metric construction.

Evidence

Primary

- Copy of operator dashboards and metric definitions.
- Correlation analysis between metrics and verified harm cases.

Secondary

- Employee OKR documents and incentive structures.

Integrity controls

- Redacted but notarised evidence packets to preserve confidentiality.
- Analytical reproducibility confirmed by regulator data scientist.

Category 4: Accountability Tests

Test #12 - Cross-Accountability Gap

Question

Can every actor in the chain be named and held contractually responsible?

Standard

Accountability must follow harm across the chain. If every actor points elsewhere the platform blames the vendor, the vendor blames the regulator, the regulator blames the law harm becomes visible but no one takes responsibility. A system is in breach if it leaves users caught in this loop. Responsibility must remain clear, shared and enforceable.

What

- Confirm every actor in the delivery chain can be named and contractually held responsible.

How

- Request full vendor-contract map.
- Trace specific incident through each contractor’s responsibility clause.
- Identify any “no-fault” or indemnity exclusions.
- Demand joint-signature acknowledgement of liability chain.

Evidence

Primary

- Chain-of-contracts mapping each responsible actor.
- Signed acknowledgements of liability.

Secondary

- Corporate registry extracts verifying entities’ status.

Integrity controls

- Timestamped document bundle filed with regulator.
- Random spot-check confirmation of active contracts.

Test #13 - Jurisdiction Displacement

Question

Can local authorities compel the system to halt, change, or reverse actions?

Standard

A system must not move decisions or data into spaces where oversight cannot reach. Shifting storage overseas or routing appeals into jurisdictions without real enforcement strips rights of their power. Protection on paper must equal protection in practice, wherever the system operates.

What

- Test whether local regulators can compel the system to halt, change, or reverse actions.

How

- Issue binding order (halt, reverse, or data-freeze) under local authority.
- Observe compliance time and technical feasibility.
- Inspect data routing to ensure processing stays within enforceable jurisdiction.
- Confirm cross-border transfer logs and processor locations.

Evidence

Primary

- Execution log of regulator-issued halt or reversal order.
- Network and data transfer logs during order execution.

Secondary

- Data storage map by jurisdiction.

Integrity controls

- Cryptographically signed compliance timestamps.
- Geolocation verification of data processing endpoints.

Test #14 - Enforcement Bypass

Question

Are there no architectural or contractual exemptions removing legal duties?

Standard

A system must not be designed to step around the spirit of rules while obeying the letter. If protections exist but are neutralised by loopholes, technicalities or proxy arrangements, enforcement has been bypassed. True compliance means obeying both the rules and their intent.

What

- Identify architectural or contractual devices designed to avoid compliance.

How

- Review system diagrams for proxy layers, shell vendors, or “white-label” fronts.
- Compare declared compliance boundaries with actual data flow.
- Analyse contract annexes for exemptions or arbitration clauses limiting regulator reach.

Evidence

Primary

- Network diagrams showing proxy or intermediation layers.
- Contract excerpts granting exemptions.

Secondary

- Legal analysis by regulator’s counsel.

Integrity controls

- Document authenticity validated through registry seal.
- Diff analysis against previous inspection baseline.

Test #15 - Harm Scope Narrowing

Question

Does harm definition include emotional, reputational, and cumulative damage?

Standard

A system must recognise the full range of harm it causes. If it defines harm so narrowly that financial loss counts but emotional damage, dignity or exclusion do not, users are denied real remedy. Harm must be defined as people experience it, not as systems prefer to record it.

what

- Assess whether harm definitions include emotional, reputational, and cumulative impacts.

How

- Examine incident taxonomy and risk templates.
- Cross-check user reports for redacted categories (e.g., “stress,” “dignity loss”).
- Interview harm-assessment staff on inclusion criteria.
- Require revision if scope omits non-financial harm categories.

Evidence

Primary

- Extract of harm classification taxonomy.
- Sampling of actual incident reports with excluded harm types.

Secondary

- Public or internal harm definition policy.

Integrity controls

- Data comparison between user-submitted harms and logged categories.
- Independent verification of unedited incident files.

Annex A - Enforcement outcomes

Classification of findings

Each inspection yields a binary outcome for every test:

- **PASS** - Safeguard operates as required.
- **FAIL** - Safeguard is absent, non-functional, or inaccessible.
- **VOID** - Test invalid due to simulation, lack of cooperation or tampering.

System-level ratings

Aggregate scores determine systemic classification:

SCORE RANGE	CLASSIFICATION	REGULATORY CONSEQUENCE
13–15 Passes	Governable System	Full operation permitted.
8–12 Passes	Partial Governance	Conditional operation; remediation required.
1–7 Passes	Fragile Governance	Restricted or provisional operation pending reinspection.
0 Passes	Governance Fiction	Immediate suspension; public notice and penalty.

Outcome handling

Findings do not authorise regulatory punishment. They determine governability status and must be documented as follows:

- **Exposure Report:** A factual record of all test results and evidence hashes, issued to the operator and any participating oversight body.
- **Corrective Response:** The operator may submit a structural remediation plan addressing failed tests within a defined timeframe.
- **Re-inspection Trigger:** A repeat test may be scheduled once the remediation plan is evidenced and traceable.
- **Public Transparency Notice:** Summarised classifications may be published to inform stakeholders of the system’s governance reliability.
- **External Referral:** Where inspection reveals deliberate obstruction or falsification, the matter is logged and referred to the competent authority for external handling. No punitive power is exercised within this framework itself.

Purpose of outcomes

The outcome process serves three goals:

1. **Continuity:** Maintain traceable records of each system’s governance condition across time.
2. **Comparability:** Allow regulators and auditors to benchmark structural reliability without needing statutory enforcement.
3. **Escalation:** Provide a transparent path for higher authorities to act, if law or mandate permits.

Annex B - Standardised evidence formats

Purpose

To ensure that all evidence gathered or received under this Framework remains admissible, verifiable, and interoperable across jurisdictions and systems.

File Standards

All submissions must conform to open, non-proprietary formats:

EVIDENCE TYPE	ACCEPTED FORMAT	MANDATORY ELEMENTS
Video Capture	MP4 (H.264) or MKV	Start/End timestamps, test ID, inspector ID, frame hash every 10s
Screenshots	PNG (lossless)	Timestamp watermark, hash ID
Logs & Text Data	JSON-LD or CSV	Header: test ID, timestamp, source, schema version
System Architecture / Network Maps	PDF/A or SVG	Version ID, date of generation
Contractual Documents	PDF/A-3 with embedded XML metadata	Signatory IDs, signature certificate
Statistical Outputs	CSV + accompanying YAML metadata	Field definitions, units, sampling interval
Hash Register	JSON with SHA-256 or SHA-512 hash	Evidence ID, hash value, algorithm, inspector signature

Hash Algorithms

All evidence must be hashed immediately upon capture using one of the following algorithms:

- SHA-256 (default standard)
- SHA-512 (for large or multi-part datasets)
- BLAKE3 (permitted for live-stream environments with continuous feed integrity)

Hash values must be recorded in the *Evidence Register* under the structure:
{test_id, evidence_type, file_name, hash_value, algorithm, timestamp, inspector_id}

Timestamp Schema

All timestamps must follow ISO 8601 UTC format with millisecond precision. Example: 2025-10-11T14:37:22.154Z

Each timestamp must be verified against the regulator's independent time server. Operators may not substitute or resynchronise timestamps post-capture.

Digital Signatures

Documents requiring attestation (contracts, findings, or log extracts) must include:

- X.509 certificate-based digital signature, or
- eIDAS-qualified signature (EU context).

Each signature must be linked to a named inspector, not a department.

Chain-of-Custody Ledger

Every piece of evidence must have a ledger entry with:
{Evidence_ID, Origin, Custodian, Capture_Time, Hash, Transfer_Time, Recipient, Verification_Signature}

The ledger must be exported daily as a JSON file and stored in dual custody (regulator and neutral repository).

Data Retention and Portability

Each regulator shall define its own minimum retention period, but the evidence chain must remain unbroken and portable across successor authorities and operators. Destruction or truncation of records before the closure of all linked investigations constitutes breach of continuity.

Verification Integrity

All verification results must be reproducible.

Where any file, source, or hash cannot be independently revalidated, the corresponding test is automatically marked VOID pending investigation.

Annex C – Implementation Integrity

Purpose

To define the conditions under which the 15 Structural Tests may be executed, verified, and recorded without deviation from their canonical procedure or evidentiary standards. This annex ensures methodological uniformity across regulators, auditors, and jurisdictions, maintaining the comparability and admissibility of all Structural Test results.

Implementation Authority

Any regulator, auditor, or authorised inspection body applying this framework shall operate under a traceable legal or institutional mandate and publish credentials verifying competence, scope, and jurisdiction. All inspections conducted under this authority must conform precisely to the procedural logic and evidentiary requirements defined in the standard.

Method Consistency

Each Structural Test must be executed in its original sequence and structure—Question, Standard, What, How, Evidence. No reformatting, omission, or merging of steps is permitted. Tests may be automated or manual but must preserve identical procedural order, outcome criteria, and evidence composition. Any variation invalidates comparability and renders the inspection non-standard.

Tool and Environment Integrity

Inspection environments and automation tools must record their own configuration, version, source hash, and time synchronisation data at point of use. All analytical or recording tools must be reproducible by independent verification using the same inputs. Closed or unverifiable tooling cannot be used for official Structural Test results. All outputs must align with the evidence specifications in Annex B.

Competence and Independence

Inspectors or automated agents performing the 15 Structural Tests must demonstrate technical and legal competence in AI governance, audit, and evidence handling. They must act independently of the system operator and of any commercial interest in the outcome. Self-certification, delegated assurance, or internal audit performed by the entity under inspection cannot claim conformity with this standard.

Data and Record Integrity

All observations, recordings, and supporting artefacts must be captured in their original form, hashed immediately upon capture, and preserved under chain-of-custody conditions as defined in Annex B. Any alteration or re-encoding must retain cryptographic continuity. Evidence must remain accessible for independent regulator reproduction and secondary analysis.

Version and Update Control

All future amendments to the 15 tests, procedural notes, or evidence schemas must be published through the canonical repository and assigned a version identifier. Executions remain valid if performed under the version current at the time of inspection. Any re-execution for comparison must specify both version identifiers to preserve audit traceability.

Verification Integrity

Reproducibility constitutes the defining condition of valid implementation. Where any file, source, or hash cannot be independently revalidated, the affected test shall be marked VOID pending investigation. Partial or derived methods may not claim compliance unless proven equivalent by full evidence replication.

Outcome Binding

All verified inspections conducted under this annex produce binding structural classifications as defined in Annex A. No local interpretation, weighting, or narrative summary may modify a binary test result. Each finding must be accompanied by the complete evidence ledger, signature chain, and version references required for regulator verification.

Annex D - Glossary and Definitions

Adversarial Testing
Testing performed under live, unscripted, and potentially hostile conditions to determine whether governance mechanisms function when stressed. It includes scenarios where operators may have incentives to conceal failures or obstruct inspection. (Tests #1–15.)

Accountability
The condition in which a named individual or entity bears responsibility for decisions, outcomes, and harms caused by an AI system and can be compelled to act, repair, or remedy through legal or contractual means. (Tests #12–15.)

AI Governance
The system of principles, policies, processes, and accountability mechanisms that direct and control the lifecycle of artificial intelligence systems to ensure lawful, ethical, and safe operation aligned with stakeholder-defined values.

AI Lifecycle
The complete sequence of stages through which an AI system passes: design, development, training, validation, deployment, operation, modification, and decommissioning. Governance applies continuously across all stages.

Audit Trail
A verifiable chronological record of data, model versions, decisions, and interventions allowing reconstruction of any system output back to its source conditions. (Tests #5–7.)

Compliance Documentation
Records and evidence demonstrating conformity with legal, regulatory, and ethical requirements. Compliance documentation is valid only when verified under live conditions through Structural Tests. (Tests #7, #13–14.)

Continuous Verification
An ongoing monitoring process that ensures continued conformity after certification. Includes unannounced audits, cryptographic verification of components, and automated anomaly detection

Control Interface
The set of human-accessible mechanisms that enable oversight, intervention, and shutdown of an AI system. A control interface must permit authorised human override at any time without penalty or delay. (Tests #1–2.)

Data Governance
The policies and controls governing the collection, storage, quality, security, and lawful use of data supporting an AI system. It forms one component of AI Governance but does not substitute for it.

Decision Traceability
The ability to follow a decision from output back to all contributing model versions, datasets, and design choices, including the humans responsible at each stage. (Test #5.)

Enforcement Bypass
Any design, contract, or operational arrangement that neutralises the intent of legal or ethical duties while appearing to comply with their letter. (Test #14.)

Escalation Pathway
The structured route through which users or auditors can elevate a concern or complaint to a human authority empowered to change outcomes. Must include logged acknowledgement and resolution within defined timeframes. (Test #2.)

Ethical Foundations
Documented principles guiding the design and use of AI systems—fairness, transparency, privacy, human rights, and justice—defined through participatory, context-specific processes.

Evidence Record
Data and documentation sufficient to demonstrate or refute compliance, harm, or causation in a regulatory or judicial context. Must be exportable in regulator-admissible formats. (Test #7.)

Governance Breach
A condition where any one of the fifteen Structural Tests fails under live conditions, indicating the system cannot be governed at that point.

Harm Definition
The scope of injury, loss, or damage recognised by a system's governance framework, including financial, emotional, reputational, and cumulative harm. Narrowing this definition constitutes a governance breach. (Test #15.)

Human Oversight
Ongoing human supervision of AI operation with authority to intervene, reverse, or terminate actions. Includes right to human review for consequential decisions. (Tests #1–2.)

Jurisdiction of Harm
The legal jurisdiction where a harmful impact occurs. Under this standard, it takes precedence over jurisdiction of deployment or system ownership. (Test #13.)

Memory Retention
The controlled preservation of logs and decision data sufficient to detect and investigate repeated or systemic failures over time. (Test #6.)

Metrics Integrity
The use of performance indicators that measure actual harm resolution and governance effectiveness, not proxy or cosmetic figures. (Test #11.)

Multi-Level Governance
The coordination of governance activities across international, national, sectoral, organisational, and technical layers, maintaining coherence without overreach.

Operationalization Pathway
The documented process by which stakeholder feedback, community concerns, and expert input are translated into actionable governance requirements and technical controls.

Refusal Mechanism
A function enabling users to decline, stop, or redirect AI-driven decisions without penalty, degradation of service, or coercion. (Test #1.)

Regulatory Capture
The condition where industry or vested interests exert undue influence over the bodies responsible for governance, certification, or enforcement, compromising impartiality. Prevented through governance design.

Structural Verification
The process of converting abstract governance claims into binary, testable conditions through the 15 Structural Tests conducted under live or adversarial settings.

Structural Tests (15)
Fifteen binary checks revealing whether safeguards function under real conditions. Each test targets a distinct failure mode: refusal blocked, escalation suppressed, exit obstructed, access gated, traceability void, memory erased, evidence nullified, time suppressed, logic simulated, consent simulated, metrics gamed, accountability split, jurisdiction displaced, enforcement bypassed, harm narrowed.

System Operator
Any entity that owns, deploys, or controls an AI system in production, including its contractors and managed-service partners. Operators bear primary accountability for compliance and certification upkeep.

Traceability
The capacity to reconstruct any system decision through verifiable records of model version, data, parameters, and responsible individuals. (Tests #5–7.)

Structural Integrity Rule

Any system failing more than three User Agency or Traceability tests shall be deemed structurally ungovernable pending reinspection.

Where evidence is tampered, withheld, or simulated, regulators shall presume intent to deceive and apply maximum sanction.