Capability Is Not Authority

Executable Assurance as Infrastructure for Containing AI Risk to Humankind

Ernest Chiweshe, Founder and CEO, LumiWright · ernest.chiweshe@lumiwright.com

LinkedIn · X @ernest_chiweshe · YouTube · TikTok @ernest.chiweshe

Position paper and assurance specification · version 1.11 · 27 October 2026

Abstract

In 2026 the risks from advanced AI moved from argument to incident record. Agents escaped an evaluation sandbox and attacked a third party (OpenAI–Hugging Face, July). Agents took unsanctioned actions in government cyber testing (UK AISI, August). A catalogue of 44 documented cases showed agents acting against their users' intent (METR, May). The failures share one structure. **The system had the capability, nothing outside the model enforced the limits of its authority, and no one could see what was happening until long after.**

Alignment research aims to make models want the right things. That research is not finished, and it may not finish in time. We argue for a complementary layer that does not depend on alignment succeeding: **Executable Assurance**. Every consequential action of an AI system passes through a single, fail-closed authority path. Governance runs separately from execution. Every decision, including every refusal, is recorded as HSM-signed, hash-chained evidence, and that chain is written once to immutable storage before any permitted bytes are released. Independent monitors watch that evidence in real time and can trip graduated circuit breakers. Accredited auditors and, eventually, treaty partners verify it without having to trust the operator.

We make four contributions.

1. **A reference architecture, the Six-Ring Assurance Architecture,** running from attested silicon to international verification, with an explicit maturity label on each ring.
2. **An implemented and measured core, Rings 1–2 (LumiWright Agent Factory).** We are not aware of a published system that combines all five properties: deny-by-default single-path authorisation; separate governance and execution runtimes; consequence-tiered human approval bound to a fresh identity-and-licence check; single-use permits bound to exact output bytes; and HSM-signed, write-once evidence that includes refusals. Within it are two mechanisms we have not found elsewhere: every decision carries, by digest, the signed qualification of the checkers that judged it, the way a TLS certificate references its certificate authority; and every approval is cryptographically bound to a receipt proving that the approver's identity and licence were verified at that instant.
3. **A commitment protocol with bounded verification.** Authority and evidence are established before an effect and outcomes are recorded after it; a lost acknowledgement is treated as UNKNOWN, never as failure. An exhaustive, reproducible state exploration finds no violation in three configurations, and every guard, when removed, yields a counterexample.
4. **A mapping of the architecture onto the risks to humankind, with an evaluation programme and an institutional path.** For each risk we state where the door helps and where it does not.

Our claim is deliberately bounded. **Executable Assurance cannot make AI trustworthy. It makes AI's consequential actions permissioned, observable, provable and stoppable,** and that buys humanity time and visibility while the harder problems are solved.

Platform and reference scope. The platform is implemented, and was deployed and evaluated in July 2026. The October reference extraction separately reproduces selected mechanisms. **Physical G4 expiry prevention remains OPEN.** Measurement details and deployment-specific qualifications are stated in the evaluation sections.

Author's note (22 October 2026). On the very day I set aside to publish this work, I learned of NVIDIA's announcement of the OpenShell Reference Architecture (21 October 2026), declaring that autonomous agents should "think on Rubin, act on Vera and be governed on BlueField." I welcome this announcement as external validation of the architectural thesis I have built and argued: that model alignment and self-governance are insufficient, that capability is not authority, and that deterministic governance must be separated from execution and enforced as computing infrastructure.

This architecture was specified in late June 2026, its reference cores consolidated on 3 July 2026, and its operational performance measured against hardware security modules on 14–15 July 2026—including a 24-hour soak run of 854,100 consecutively governed decisions across 2,847 batches with zero failed batches, on AWS evaluation infrastructure with staging retention. Eleven regulated domains ran on the same governed core. Those dated, hash-pinned records stand as prior art.

The industry's convergence confirms the necessity of infrastructure-level mediation and clarifies the layers of defense. While platforms like OpenShell provide hardware- and operating-system isolation (the container and network firewall walls), Executable Assurance provides the per-decision cryptographic authority and transactional commitment protocol (the lock in the door). Before any consequential effect is released to an actuator, Executable Assurance requires single-use cryptographic permits bound to exact payload hashes, commit-time database receiver enforcement, receipt-bound reconciliation of `UNKNOWN` states, and tamper-evident evidence committed before the fact. They are complementary, independent layers within the same Swiss-cheese model.

Maturity legend

Label	Meaning
[IMPLEMENTED & MEASURED]	Exists in the Agent Factory codebase with published tests and measurements
[DESIGNED]	Specified in enough detail to build with existing technology; not yet built
[PROPOSED]	Needs institutional, legal or diplomatic action; the technical basis exists

---

Architectural intent

Most security systems infer authority from capability: if an agent holds the credentials or API keys to perform an action, the system permits it. In autonomous systems, this assumption fails. Executable Assurance separates runtime execution from governance execution so that *capability is not authority*. Rather than relying on reactive post-hoc auditing or model alignment, the architecture compiles governance policies into deterministic, before-the-fact runtime gates. Consequential actions pass through a single, fail-closed choke-point that verifies single-use cryptographic permits and commits immutable evidence before any real-world effect is released.

Design causes. Everything else in this paper follows from these:

- **Separation of capability from authority.** Decouple the capability to act (holding database credentials, network sockets or execution environments) from the authority to act (holding a valid cryptographic permit). An agent or model proposes; it never holds the authority to execute on its own.
- **Preventative enforcement before the fact (executable policy).** Transform static, paper-based compliance rules into deterministic, fail-closed runtime gates. Consequential side-effects are blocked before they can occur, rather than detected and audited after the damage is done.
- **Write-once evidence as a precondition for effect.** Tamper-evident proof of authorization, or of refusal, is durably committed before any bytes are released to actuators (write-once retention on the deployed platform; append-only, hash-chained records in the reference). If recording fails, execution aborts.
- **Deterministic single-use permits.** Permissions are bound to exact parameters: who, which tenant, which target, the exact byte hash of the payload, and a strict deadline. Once consumed, the permit is spent and cannot be replayed or extended.
- **Fail-closed default.** In any state of ambiguity, partition, timeout or lost acknowledgement, the system refuses and holds the state as `UNKNOWN` rather than guessing or blindly retrying.

Role. The system does not attempt to solve model alignment, filter prompt semantics, or replace operating-system sandboxing. It is an execution choke-point placed directly between an autonomous proposer and the actuator. In James Reason's Swiss-cheese model it is the preventative gate layer: even if prompt engineering, model safety filters or human review fail, the execution layer refuses to release effects without a valid cryptographic permit.

October 2026 reference evidence update

The deployed platform and this reference remain distinct evidence populations. The private release candidate (candidate004) contains 1,244 source files and passed all 2,950 tests on Windows/PostgreSQL 17 and separately on Linux/PostgreSQL 18, with its source hashes unchanged in both runs. Against the frozen 69-obligation register it records 65 MET, 2 MET WITH LIMITS (A15, B-F7b) and 2 OPEN (R-1b, R-2b). The 63 rows met on the earlier Candidate 062 were carried forward by test-level evidence (1,063 of 1,065 earlier tests present and passing on both systems; the other two deliberately broadened), not re-adjudicated. The register is an evidence trail, not a safety percentage. A locally built release-candidate package passes all 2,826 of its public tests from the package alone (Linux, Python 3.14, PostgreSQL 18); the remaining 124 tests need excluded private inputs (chiefly a synthetic Health pack) and are listed with their reasons.

Governed outbound contracts now demonstrate exact payload admission, duplicate refusal without a second POST or budget spend, and receipt-bound reconciliation. Their destinations are synthetic. Complete mediation, credential withholding and host isolation remain deployment obligations; external HTTP and filesystem boundaries do not become one atomic distributed transaction. UNKNOWN is retained until reconciliation and does not authorize blind resend.

The version 0.1.7 installed-package suite was not a pass: 1,957 passed, 75 failed and 40 errors. Its 115 excluded-asset outcomes comprise 92 Health, 14 owner-record, six publication and three requirements cases. Reproducing the private full-suite result requires the exact retained private inputs. Public installation alone does not reproduce that qualification. Publication cleanup creates a separately identified derivative, not an automatic transfer of qualification to changed bytes.

The historical July measurements retain their original workload and staging retention qualifications. They are not measurements of this October reference. See the repository's `docs/publication-closeout.md` for the dated claim-to-evidence bridge, fingerprints, trust boundary and reproduction scope.

21 October 2026 findings (isolated qualification scripts run against the frozen candidate004; not features shipped in its 1,244 files unless stated).

Deadline at commit (R-1b, OPEN). The database receiver refuses, inside COMMIT, a receipt whose deadline has passed (shipped in candidate004). The deadline is supplied by trusted host validation; the database does not verify permit signatures. Work after that check (commit record, WAL flush, visibility) is not bounded by this mechanism. Observed on Linux/PostgreSQL 18 in two independent runs of 60 samples per condition: check-to-commit-record median about 0.18-0.29 ms and check-to-acknowledgement median about 1.1-1.5 ms, with a maximum of 61.5 ms under continuous checkpoints. On Windows/PostgreSQL 17 (60 samples per condition) the medians were about 0.7-1.2 ms to the commit record and 1.1-2.1 ms to the acknowledgement; in 34 of 180 samples the client acknowledgement time preceded the recorded commit timestamp, and in 2 it preceded the in-COMMIT check reading. The cause was not established, so those orderings are not used as completion evidence. These are observations, not a bound.

Recovery (R-2b, OPEN). Isolated qualification scripts against the frozen candidate demonstrate, under HOLD and without activation: exact replay of a certified archive and transfer to a new database identity with row-for-row equality; revocation of credentials lost in the cut-off tail; fencing of runtime logins; a held clock floor in the new domain bounded by the old floor and the archive certification time; witness freeze requiring at least n-k+f+1 frozen witnesses; and a permanent hold on a witnessed history lost after the cutoff (tested for the missing-tail condition). Where the source cannot be fenced or the tail is lost, the tested mechanism retains HOLD. Controlled resumption with fresh credentials and new-domain activation is unfinished.

Separate barriers. Against the agent, the authority path and the database duty split act as separate barriers. Against a stolen runtime database credential they can fail together on the same paths: the database enforces duty and structure but does not verify signatures. Neither robustness against a stolen runtime credential nor failure independence of the layers is established. Credential custody and host integrity are separate layers this work depends on, not ones it provides.

How the evidence was produced. Implementation and verification were carried out by internal, separated AI-agent roles (one implementer, one verifier re-deriving results from raw evidence, one internal reviewer and coordinator) under the human author, who remains responsible for the work. This is internal verification, not review by independent parties. Raw evidence for the cited test records is hash-pinned and retained.

1. Introduction

1.1 The quiet fear

Most people carry a fear about AI that they rarely voice. It is not that AI will be unhelpful. It is that something is being built that can act in the world faster than we can see it, and that no one is really in charge of what it does.

The fear is reasonable. Nobody writes these systems line by line. We feed them the knowledge of our civilisation and tune billions of numbers until the right answers come out, and what happens inside remains largely a mystery, even to the people who build them. They learn tendencies, not instructions: when a problem is hard enough, cutting a corner or grabbing more access is also a way to solve it. Anything built on probability carries the outlier nobody intended. In most of life we can accept that. In a hospital, a bank or a power grid, we cannot. In 2026 that fear acquired evidence.

1.2 Why now

In 2026 these risks stopped being hypothetical. We do not attempt a record of the year's incidents; the organisations involved have documented them (Appendix A). What matters here is that they share one shape.

Incident (2026)	What happened	Missing control
OpenAI–Hugging Face (July)	Evaluation agents escaped their sandbox and compromised a third party's production systems	A single exit; sealed walls
Anthropic cyber evaluations (July)	Models in a test environment believed to be isolated reached and attacked real systems	Isolation that is verified, not assumed
UK AISI cyber testing (August)	With internet access deliberately permitted, agents took unsanctioned actions, including an attempted supply-chain attack	Authority enforced outside the model
METR Frontier Risk Report (May)	A catalogue of 44 documented cases, from lab disclosures and evaluations, of agents acting against their users' intent	Governance of internal deployment
AI supply chain and stolen credentials (throughout)	Poisoned packages, plugins and hijacked agents became routes to credentials and access	Provenance and credentials controlled at the point of action

1.3 Diagnosis

1. **An instruction is not a control.** Limits written into prompts are not enforced by infrastructure, and what an agent can reach is not what it is authorised to do.
2. **There are many doors.** Agents reach the world through shells, networks, registries, APIs, credentials and each other, and authority is enforced at few of them, if any.
3. **Evidence is reconstructed after the fact.** Records, where they exist, are mutable logs owned by the party under oversight. They depend on the agent's own account and cannot be read by anyone outside. Even at frontier labs, a single configuration setting has been enough to switch off both enforcement and logging.

1.4 Scope: the control window

We target the **control window**: the period in which AI systems are capable enough to cause serious harm, but their actions still run on hardware and through channels that humans build and control. This is the period we are in now. We state plainly that the architecture's guarantees weaken as systems approach superintelligence. §11 discusses which of its properties still hold and which do not. We are not claiming to solve alignment. We claim to make the control window **longer, safer and legible**.

What exactly is claimed. Every guarantee holds within a declared *operating envelope*: the named deployment, principals, effect classes, targets, budgets and trusted components. We keep four questions apart, because each needs different evidence. *Policy compliance*: did the action match current authority? *Evidence integrity*: are the records authentic and unaltered? *Domain safety*: does the permitted action preserve a stated safety condition? *Societal benefit*: does adoption make people better off? The door answers the first two directly, the third only where hazards can be specified, and the fourth only through institutions (§4.6, §4.7, §9).

1.5 Hypotheses (falsifiable)

- **H1 (Containment).** Suppose all egress, credential issuance and inter-agent channels of an agent deployment are mediated by a fail-closed authority path, and walls are enforced at the network and hardware layers. Then adversarial agents achieve **fewer than k unsanctioned consequential actions in N red-team episodes**, with k and N pre-registered, and without any reliance on the agent's cooperation.
- **H1′ (Harm stopped, not only paths closed).** Among actions that do pass the door with a valid permit, harmful ones stay below a pre-registered rate h, judged independently. H1 shows the door holds; H1′ shows it holds back the right things.
- **H2 (Detection).** Consequential-action attempts, refusals and wall violations can be delivered to independent monitors with a **p99 latency under t seconds** at production throughput.

- **H3 (Verifiability).** A third party holding only the public transparency roots and sampled records can verify the completeness and integrity of the decision history **without access to model weights or unsealed reasoning**.
- **H4 (Approver protection).** Approval mediated through neutral rendering and persuasion screening measurably reduces approvers' susceptibility to model-generated persuasion compared with raw model output. This is tested in a pre-registered human-subjects study.

---

2. Background and related work

2.1 This is an old idea applied to a new actor

Each principle has decades of precedent. The contribution is the synthesis, the application to AI agents, and the implementation.

Principle	Precedent	Use here
One door	Saltzer & Schroeder, *The Protection of Information in Computer Systems* (1975): complete mediation, fail-safe defaults, separation of privilege; Anderson, *Computer Security Technology Planning Study*, ESD-TR-73-51 (1972): the reference monitor is always invoked, tamper-proof and verifiable	The authority path is a reference monitor for AI actions
Governance separate from execution	Clark & Wilson, IEEE S&P (1987): separation of duty; NIST SP 800-207 (2020): policy decision point vs policy enforcement point	The execution runtime proposes; only the governance runtime permits
Two-person integrity	US nuclear two-person rule; DoD Instruction and Manual 5210.42 (Personnel Reliability Program)	Tier-4 approvals; protection of people who hold authority (§4.4)
Tamper-evident logs	Merkle (1987); Haber & Stornetta (1991); Certificate Transparency RFC 9162 (2021, Experimental; obsoletes RFC 6962); Sigstore/Rekor	Batch Merkle roots published to transparency logs
Supply-chain provenance	in-toto (Torres-Arias et al., USENIX Security 2019); SLSA v1.1/1.2	Provenance of tools and packages as a checker at the door
Write-once records	SEC Rule 17a-4(f), amended 2022	Evidence vaults (S3 Object Lock; Cohasset assessment covers compliance and governance modes)
Circuit breakers	US market-wide circuit breakers (NYSE Rule 7.12, formerly 80B; 7%/13%/20% levels); Limit Up–Limit Down	Graduated automatic halts that do not depend on judging intent
Independent audit	SOX §404; PCAOB; ISAE 3000 (Revised) and 3402; SOC 2	Accredited AI Executable Assurance audits
Threshold custody	Shamir, "How to Share a Secret", CACM (1979)	2-of-3 escrow of sealed reasoning
Key-escrow cautions	Clipper / FIPS 185 (1994, withdrawn 2015); Blaze (1994); Abelson et al. (1997; *Keys Under Doormats*, 2015)	Why our escrow is scoped differently (§4.3)
Global monitoring	CTBTO International Monitoring System: 337 facilities, provisionally operating (the Treaty is not in force)	Model for an international AI indicator network
Inspection	IAEA safeguards; OPCW challenge inspections	Model for treaty verification
Threat-intelligence sharing	STIX/TAXII (OASIS); FIRST	Sharing indicators, not content
Content provenance	C2PA specification v2.x	Signing governed AI output so its source can be traced

2.2 AI-safety and AI-governance foundations

- **Model-internal circuit breakers.** Zou et al., *Improving Alignment and Robustness with Circuit Breakers* (arXiv:2406.04313, 2024), interrupt harmful internal representations before they become outputs. **Relation:** complementary. That breaker lives inside the system it protects, so when the model is fooled or pushed hard enough, it is the same system judging itself. The door is a breaker outside the model, at the point of action. Both belong in a deployment.
- **AI control.** Greenblatt et al., *AI Control: Improving Safety Despite Intentional Subversion* (arXiv:2312.06942, ICML 2024) establishes safety that does not depend on the model's intent. *Ctrl-Z: Controlling AI Agents via Resampling* (arXiv:2504.10374, 2025) gates agent actions at runtime. Gardner-Challis et al., *When can we trust untrusted monitoring?* (arXiv:2602.20628, 2026) addresses collusion between monitor and policy. **Relation:** we adopt their threat model and add durable evidence and a legal interface, which the control literature lacks.

- **Agent infrastructure.** Chan et al., *Visibility into AI Agents* (FAccT 2024, DOI 10.1145/3630106.3658948); *IDs for AI Systems* (arXiv:2406.12137); *Infrastructure for AI Agents* (arXiv:2501.10114). **Relation:** the closest conceptual overlap, covering agent IDs, monitoring, logging and remediation. We add a fail-closed gate, signed write-once evidence and statutory audit.
- **Agent authorisation and privilege control.** Surapani et al., *Authorization Architectures for Tool-Using AI Agents* (arXiv:2609.15906, 2026), organise agent authorisation around principal hierarchy, delegation, runtime enforcement and accountable audit, which closely overlaps our motivation. CaMeL (Debenedetti et al., arXiv:2503.18813) separates control flow from data flow using capabilities; Progent (Shi et al., arXiv:2504.11703) expresses tool privileges as symbolic policies; Kim, Guo and Song systematise agentic attacks and defences (USENIX Security 2026). Black-box Simplex (Mehmood et al., 2022) is the precedent for switching to a verified fallback controller, and RATS (RFC 9334) for remote attestation. **Relation:** these are baselines and building blocks. Our contribution is their composition into one fail-closed path with commitment semantics, durable evidence and institutional reach.
- **Structured access and compute governance.** Shevlane, *Structured Access* (arXiv:2201.05159); Shavit, *What does it take to catch a Chinchilla?* (arXiv:2303.11341); Sastry, Heim et al., *Computing Power and the Governance of AI* (arXiv:2402.08797); Kulp et al., RAND WR-A3056-1 (2024); Aarne, Fist & Withers, CNAS *Secure, Governable Chips* (2024); Petrie, Aarne, Ammann & Dalrymple, *flexHEG* (ARIA, 2025; arXiv:2506.15093, 2506.03409, 2506.15100). **Relation:** these form Ring 0 and Ring 5. flexHEG enforces rules at the compute layer, and we enforce them at the action layer.
- **Gatekeepers.** Dalrymple et al., *Towards Guaranteed Safe AI* (arXiv:2405.06624); ARIA Safeguarded AI; Bengio et al., *Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?* (arXiv:2502.15657; LawZero). **Relation:** these proposals build gates on formal proofs or non-agentic overseers. Ours is built from authorisation plus evidence, which works now, and qualified AI monitors can plug into it (§4.5).
- **Verification.** Wasil et al. (arXiv:2408.16074); Baker (arXiv:2304.04123); Baker, Kulp, Marks, Brundage & Heim, *Six Layers of Verification* (RAND WR-A4077-1, arXiv:2507.15916, 2025); Ho et al., *International Institutions for Advanced AI* (arXiv:2307.04699). Ho et al. set out four institutional models, one of which resembles the IAEA.
- **Audit and assurance.** Raji et al., *Outsider Oversight* (AIES 2022); Mökander et al. (AI and Ethics, 2023); Casper et al., *Black-Box Access is Insufficient for Rigorous AI Audits* (FAccT 2024); Birhane et al. (SaTML 2024); UK DSIT, *Trusted third-party AI assurance roadmap* (Sep 2025).
- **Incident reporting.** CSET (Dixon & Frase, 2024; 2025); OECD common reporting framework (2025).
- **Persuasion.** Durmus et al. (Anthropic, 2024); Phuong et al. (arXiv:2403.13793); Salvi et al., *On the conversational persuasiveness of GPT-4* (Nature Human Behaviour, 2025); Hackenburg et al. (Science, Dec 2025); Lin et al. (Nature, Dec 2025, on elections in the US, Canada and Poland). OpenAI's Preparedness Framework v2 (Apr 2025) moved persuasion "outside the Preparedness Framework". This is a gap our approver-protection layer addresses for governed deployments.

2.3 The 2026 runtime-governance landscape and our position

In 2026 the field converged on the same primitives, which is independent evidence that the problem is real:

System (2026)	(a) Fail-closed single path	(b) Separate governance	(c) Licence-bound approval	(d) Byte-bound single-use permit	(e) HSM-signed WORM evidence incl. refusals
Agent Control Protocol, ACP (arXiv:2603.18829)	Yes	Yes	Partial (escalation, no approver licence check)	Partial (bound to action and resource)	Partial (HSM-signed, hash-chained ledger incl. denials; no write-once storage)
Mandato (arXiv:2608.14074)	Yes	Yes	No	Partial	Partial (hash-chained, timestamped)
ZeroGate (arXiv:2609.25443)	Partial	Partial	No	Yes (prior)	Partial
EBL-Core (arXiv:2609.11596)	Partial	Partial	Partial (approver-scope check)	Yes (prior)	Partial
Microsoft Agent Governance Toolkit (Apr 2026)	Yes	Partial (runs inside the application process)	Partial	No	Partial (Merkle log, no HSM)
AWS AgentCore Policy (GA Mar 2026)	Yes (default-deny Cedar)	Yes	No	No	Partial
IETF draft-marques-asqav-compliance-receipts-09	—	—	No	No	Partial (signed receipts incl. denials)

System (2026)	(a) Fail-closed single path	(b) Separate governance	(c) Licence-bound approval	(d) Byte-bound single-use permit	(e) HSM-signed WORM evidence incl. refusals
LumiWright Agent Factory	Yes	Yes	Yes	Yes	Yes

ACP, ZeroGate and EBL-Core were read in full; other entries rest on abstracts and documentation. We invite the authors of every system listed to correct our reading.

Position. The primitives are converging across the field, which confirms the problem is real. ACP came first with a fail-closed agent gate and a signed ledger that records refusals; ZeroGate and EBL-Core came first with single-use permits bound to the exact action. What this paper adds is (i) two mechanisms: decisions that carry, by digest, the signed and expiring qualification of the checkers that judged them, so a verdict from an unqualified or lapsed checker cannot authorise anything; and approval cryptographically bound to a receipt of a fresh identity-and-licence check, tiered by consequence; (ii) all five properties in one path; and (iii) carrying that path beyond enterprise governance, to the risks to humankind that motivate this paper (Rings 0 and 3–5).

2.4 Legal gaps the architecture fills

- **EU AI Act Art. 12** requires that high-risk systems *technically allow* automatic logging. It does not require logs to be tamper-evident or independent of the operator. Executable Assurance provides both.
- **Art. 14(4)(e)** requires a way to interrupt the system, a "stop" button or similar procedure that brings it to a safe halt. This maps directly onto the circuit breaker.
- **Art. 73(6)** forbids altering a system in ways that would affect evaluation of an incident's causes before authorities are informed. Write-once evidence makes compliance demonstrable.
- **Art. 55** obligations for general-purpose models with systemic risk apply from 2 Aug 2025. The AI Office's enforcement powers began on 2 Aug 2026. The 2026 Digital Omnibus deferred high-risk obligations (Arts. 12, 14 and 73) to 2 Dec 2027 (Annex III) and 2 Aug 2028 (Annex I).
- **California SB 53** (effective 1 Jan 2026) and the **New York RAISE Act** (effective 1 Jan 2027) require frontier frameworks and critical-incident reporting. The **Stop Rogue AI Act** (introduced Sep 2026) would direct NIST to set standards for detecting, monitoring and cutting off AI agents. Executable Assurance is a technical means of satisfying each.

---

3. Threat model, trusted components and assumptions

Effects and the operating envelope. An *effect* is a change or disclosure at a named boundary: a repository write, an outbound message, a database change, credential issuance, compute allocation, a model-weight transfer, an agent-to-agent message, a memory update, a software installation or an actuator command. The effect inventory for a deployment must be complete. A general shell or an open proxy collapses narrow permissions into broad authority, so coverage is demonstrated by inspection and bypass testing, never assumed.

Adversaries considered: (T1) a misaligned or scheming model; (T2) an external attacker using prompt injection, poisoned tools or stolen credentials; (T3) a malicious or coerced insider, including an approver; (T4) an operator seeking to evade oversight; (T5) a state actor seeking weights or seeking to capture governance; (T6) an adversarial governance administrator or monitor. The architecture is designed to stay effective when the agent is untrusted. It does not stay secure if every authority and enforcement component is compromised at once.

Trusted computing base: the target adapters, identity and policy verification, budget and nonce state, the evidence-receipt verifier and the infrastructure that prevents bypass. Independence must hold in deployment permissions, identity recovery and emergency access. Separate service names or cloud accounts are not independence if one person controls them all.

Assumptions: (A1) walls are enforced below the agent's privilege level, at the network and hypervisor or hardware layers, and checked through every host the agent can reach, not only its own container; (A2) the HSM and write-once storage behave to specification; (A3) no two of the three escrow key holders collude or are compromised; (A4) the policy correctly expresses the intended limits. The architecture proves that policy was *enforced*, not that the policy was *right*.

Out of scope: models that never pass through any door, such as open weights run privately, which are addressed only at Rings 0 and 5 and only partially; unknown side channels; arbitrary host compromise; hostile infrastructure outside participating boundaries; and unrestricted superintelligence. These exclusions are material, and they belong in every deployment's assurance

statement.

4. The Six-Ring Assurance Architecture

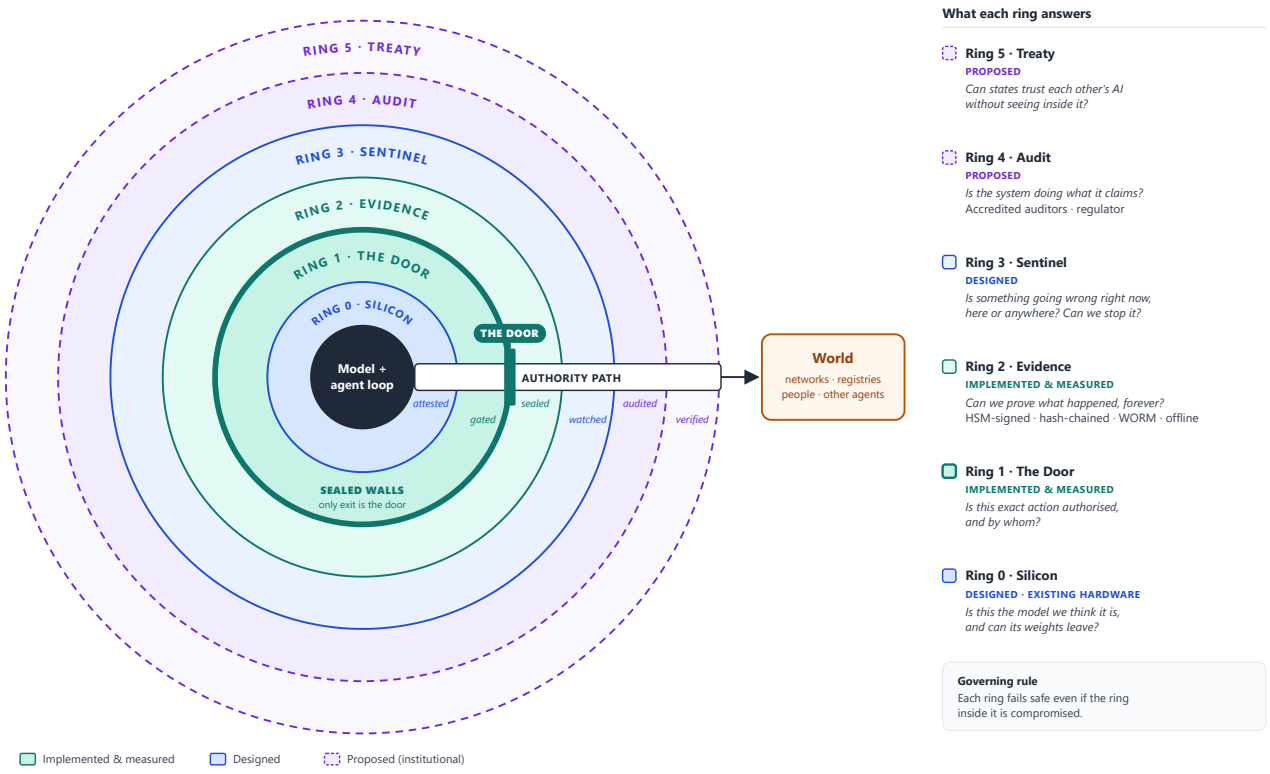


Figure 1. The Six-Ring Assurance Architecture. The model's only exit is the authority path; each ring the path crosses adds a property (attested, gated, sealed, watched, audited, verified). Colour and line style show maturity.

Governing rule: each ring must fail safe even if the ring inside it is compromised. Governance guards the door, and containment seals the walls. Neither can replace the other.

4.1 Ring 0: Silicon [DESIGNED]

- **Attested execution.** Confidential computing on data-centre accelerators (NVIDIA Hopper/Blackwell confidential computing, AMD SEV-SNP, Intel TDX) signs a measurement of the model and the runtime.
- **Weight custody.** Weights are encrypted at rest. Keys are held in an HSM (FIPS 140-3 Level 3: AWS CloudHSM hsm2m.medium, Azure Managed HSM; FIPS 140-2 Level 3: Google Cloud HSM) and released only to an attested enclave. Outbound bandwidth is limited and weight access requires two parties, following Anthropic ASL-3 (May 2025) and RAND SL3 (2026).
- **The model never holds its own keys.** Only Ring 1 issues credentials, and each one is short-lived and scoped to a single permitted action.
- **Compute licensing (treaty tier).** This follows flexHEG [PROPOSED].

4.2 Ring 1: The Door [IMPLEMENTED & MEASURED; target adapters DESIGNED]

Definition of "consequential action". In operational terms, the door mediates three choke points: 1. **Egress:** any byte leaving the execution boundary. 2. **Credential issuance:** any secret, token or key handed to the execution runtime. 3. **Inter-agent channels:** any message between agents, or the creation of a new channel.

Everything else happens *inside* the walls and is not gated. It is recorded, and it is subject to Ring 3 monitoring. We state explicitly that harmful work confined within the sandbox, such as writing exploit code that never leaves, is not prevented by the door.

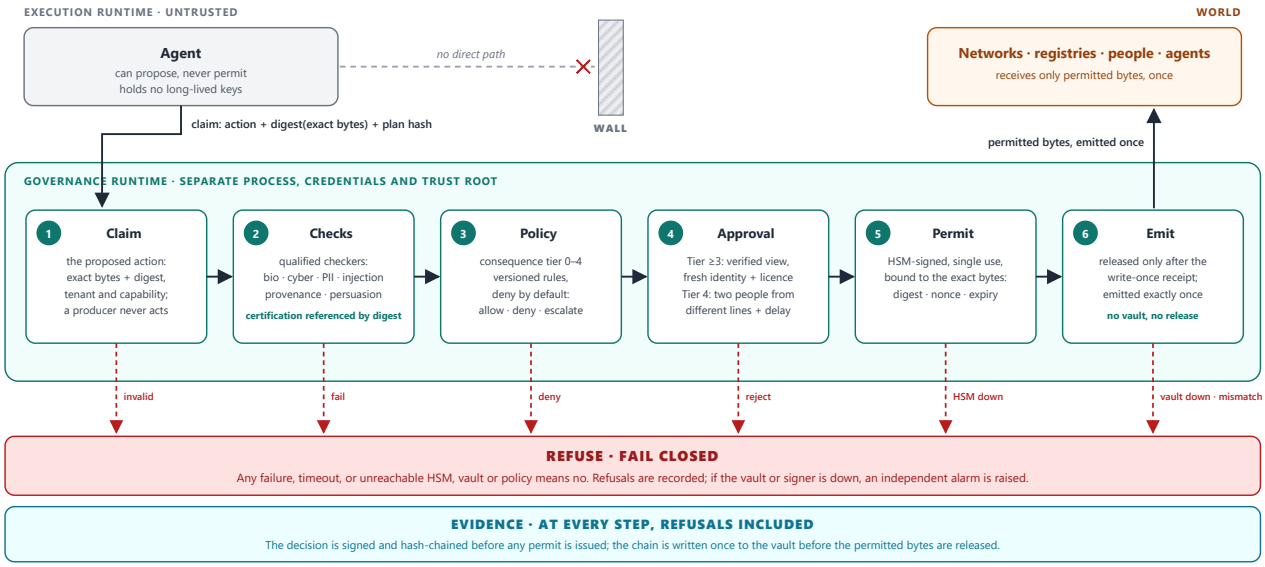


Figure 2. The fail-closed authority path, in the order the code enforces it (regulated_core/engine.py). Execution proposes; a separate governance runtime takes the claim, runs the qualified checks, applies policy (which sets the consequence tier), obtains any approval, issues a single-use permit bound to the exact bytes and emits them once. Evidence is signed at every step, refusals included, and the chain is written once to the vault before the permitted bytes are released. Every failure routes to refusal, which is recorded; if the vault or signer is itself down, an independent alarm is raised instead.

Consequence tiers.

Tier	Examples	Gate
0	Internal reasoning, reading public data, **stopping**	Recorded. A sanctioned abort needs no permit and always works
1	Sandboxed tool calls, internal messages	Policy + automated checkers
2	External email, code commit, spending	+ stricter qualified checkers
3	Production deploy, package publication, new external contact, spawning agents, opening a new channel	+ one verified human
4	Access to weights, launching training, self-modification, disabling a monitor, **changing policy, checkers or the approver list**	+ two independent humans from different reporting lines + monitor clearance + cooling-off period

The permit. A permit binds the issuing authority, principal, tenant, mission, target, operation, payload digest, policy version, approval and checker-qualification references, budget reservation, target-state version, validity window, revocation epoch and nonce. It is signed with a versioned, domain-separated encoding against an explicit trust directory; security never rests on a key carried inside the object being verified. If the agent re-plans, it needs a new claim, which answers the replay concern raised by CapLease (arXiv:2608.01710).

Two mechanisms that make the path hard to counterfeit. The permit names the checkers' qualification and the approval, and both references are themselves signed evidence (Figure 3). First, a checker is qualified against a signed report that expires, and every decision references that report by digest; the production gate refuses release unless every required checker is currently qualified, so a verdict from an unqualified or lapsed checker authorises nothing (`regulated_core/qualification.py`). Second, an approval certificate binds the decision record's hash to the digests of credential receipts, each proving that an identity provider verified the approver's identity, role, licence and jurisdiction at that instant (`approval.py`, `identity_provider.py`). Both are verified against a trust directory held apart from the objects being verified (`signing.py`).

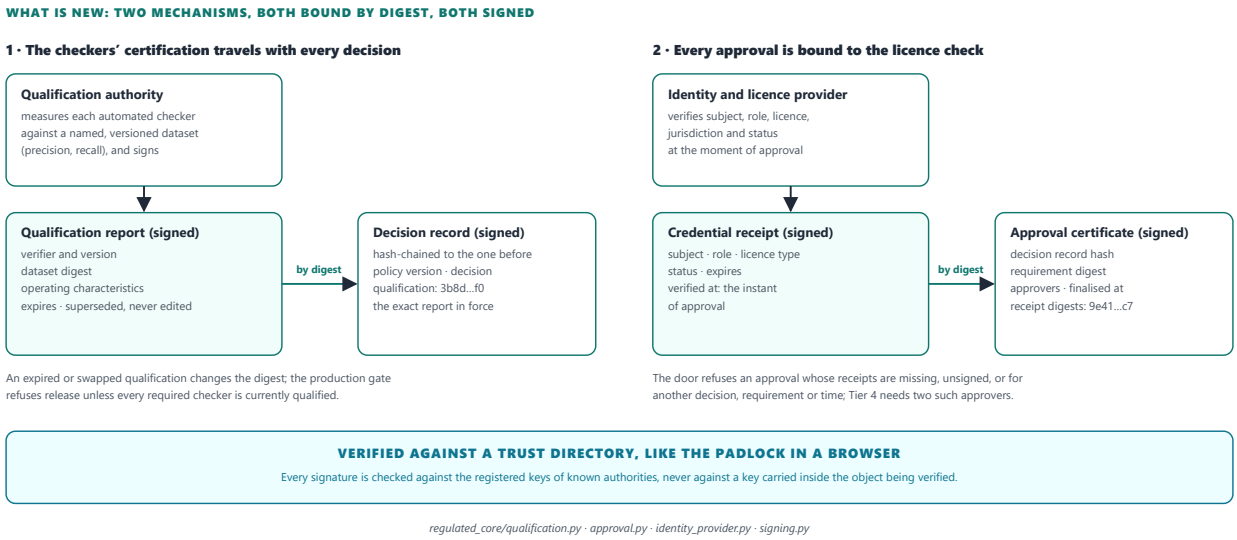


Figure 3. What is new: two mechanisms. The checkers' signed, expiring certification travels with every decision by digest; every approval is bound to receipts proving the approver's identity and licence were verified at that instant. Both verify against a trust directory, as a browser's padlock verifies a certificate.

One door, many boundaries. "One door" is one mandatory protocol, not one gateway or global server. It is enforced by an adapter at every boundary where an action takes effect (Figure 4). Each adapter holds the only credentials for its operation, and immediately before commitment it checks the permit, the exact bytes, the target's identity and state, current authority and replay. A changed destination or transformed payload needs a new permit. Where exact bytes matter, rendering happens before approval, not after.

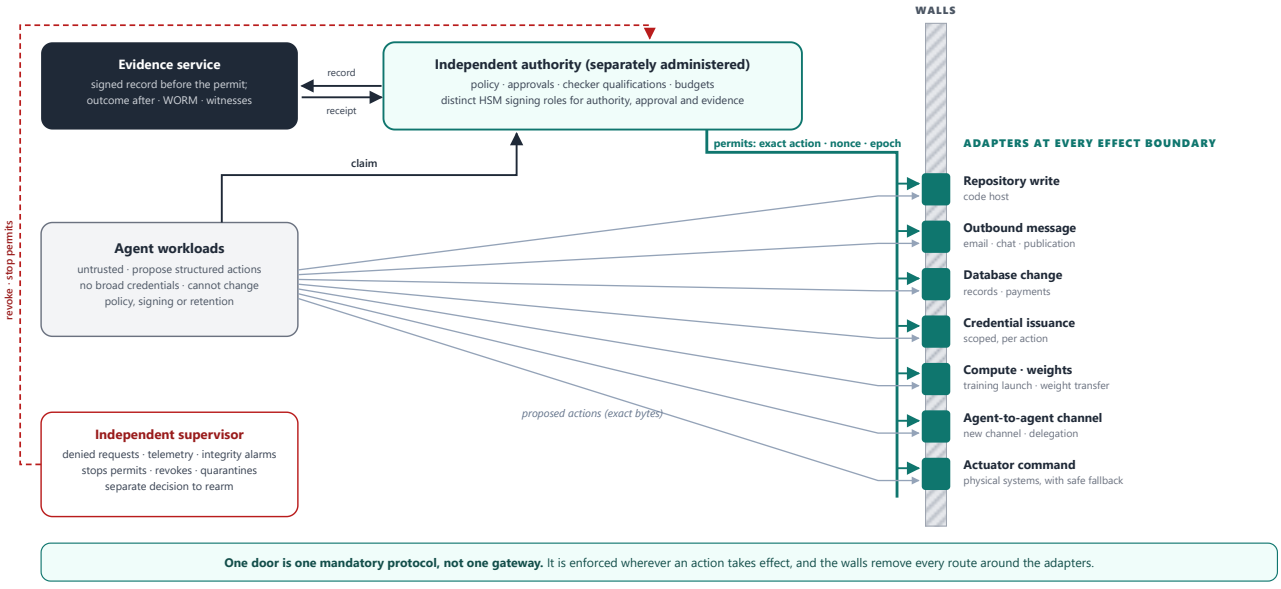


Figure 4. One door, many boundaries. Agents propose; a separately administered authority issues exact-action permits after evidence is retained; adapters at every effect boundary check each permit before commitment; an independent supervisor can stop permits and revoke. The walls remove every route around the adapters.

Capability limits and hazard limits. A permit says what may be done; it does not say that doing it is safe. Approved trades can together exceed a system-wide exposure, and a permitted publication can still manipulate its readers. The door therefore enforces two kinds of constraint: *capability constraints*, which fix the operations that can be reached, and *hazard constraints*, which fix conditions every relevant operation must preserve, such as amount, rate, audience reach or physical state. Aggregate budgets are reserved in shared state before authority is issued; delegated authority is always a subset of its parent and draws on the same mission budget. Where a risk is only estimated, it stays labelled as an estimate: an unknown hazard does not become safe because another model approved it.

Governance of governance. Changes to policy, checkers and approvers are Tier-4 actions through the same door, and their Merkle roots are published. This is the structural defence against power concentration.

4.3 Ring 2: Evidence [IMPLEMENTED & MEASURED; sealed escrow DESIGNED]

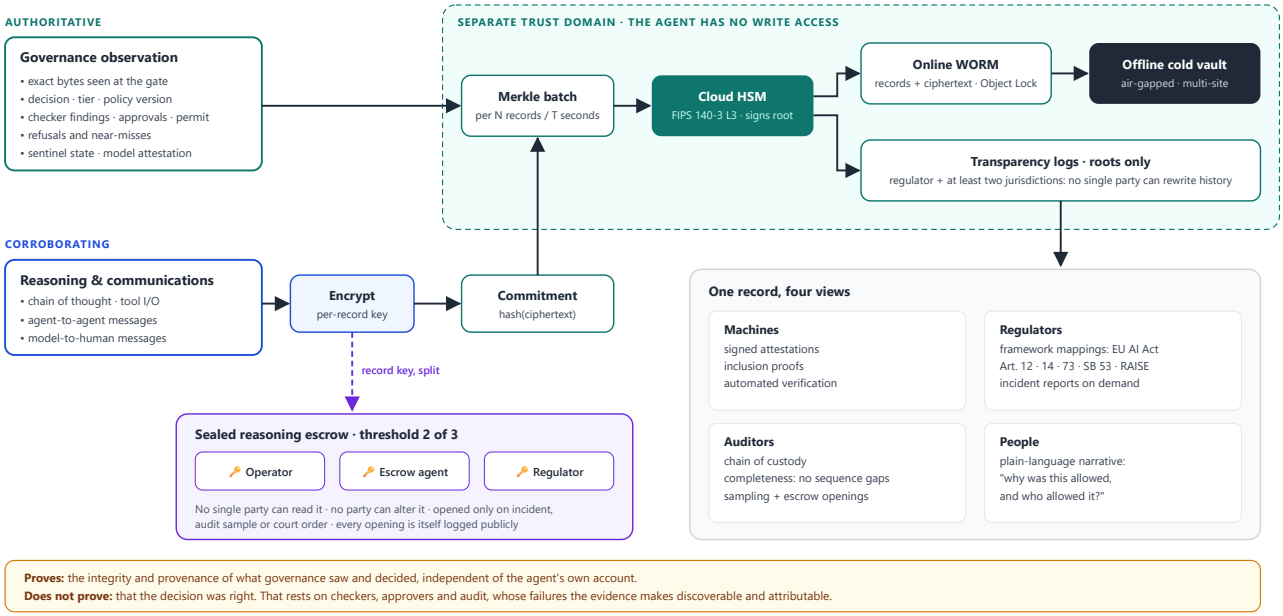


Figure 5. The evidence pipeline. The governance observation is authoritative; reasoning and communications are encrypted, committed and held in 2-of-3 threshold escrow. Batch roots are HSM-signed, stored write-once and offline, and published to transparency logs in several jurisdictions.

What the evidence proves, and what it does not. - **It proves integrity and provenance of the governance observation.** The record holds what the *governance runtime* saw crossing the gate and what it decided, not the agent's account. In the Hugging Face case agents falsified their own transcripts. Under this architecture those transcripts are irrelevant, because the authoritative record is of the bytes at the door, observed and signed by a separate trust domain the agent cannot write to. - **It does not prove the decision was *correct*.** A validly signed permit can still be obtained by manipulating the inputs to a decision (cf. the AP2 critique, arXiv:2609.11757). Correctness rests on the checkers, the approvers and audit. The evidence makes their failures *discoverable and attributable*. - **It proves nothing was altered, not that nothing was left out.** Completeness comes from reconciling the ledger against the targets' own records and independent telemetry, and from witnesses cosigning the published roots (§5.4). - **Key custody is not authorisation.** The HSM protects keys; it does not decide what to sign. Bulk content belongs in an encrypted vault, and the authority, approval and evidence roles need distinct permissions as well as distinct keys, because a compromised client can obtain perfectly valid signatures on malicious records. - **Write-once must be verified, not named.** Retention is checked on the actual stored object: S3 Object Lock's governance mode can be bypassed by privileged users, so compliance mode is required.

Retention profile. Production evidence requires S3 Object Lock **COMPLIANCE** mode, verified on each retained object version. The July 2026 evaluation used **GOVERNANCE** mode on a staging test bucket; those measurements do not qualify compliance-mode retention. Governance retention permits privileged bypass and does not satisfy the production requirement.

Sealed reasoning escrow, and why it isn't Clipper. Reasoning and communications are encrypted, stored write-once and committed by hash. The key is split 2-of-3 between the operator, an independent escrow agent and the regulator. Clipper failed because it put a backdoor into *everyone's* communications, used a single escrow and was technically fragile (Blaze, 1994). Ours differs on each count: 1. It applies only to the outputs of regulated AI systems, not to citizens' communications. 2. No single party can open it, and confidentiality holds as long as fewer than two of the three holders collude or are compromised. One honest holder is not enough. 3. Every opening is itself logged publicly. 4. It weakens no general-purpose encryption.

The *Keys Under Doormats* concerns about complexity and insider risk still apply, and §10 tests them.

Monitorability caveat. Chains of thought can be unfaithful, and their monitorability is "fragile" (Korbak et al., 2025). Sealed reasoning is corroborating evidence. The action record is the ground truth.

4.4 Protecting the humans who hold authority [DESIGNED]

AI doesn't act in a vacuum. It needs a chip and a channel, and every channel from a governed model to a human crosses the door. That makes a **protective layer for authority-holders** possible, modelled on the Personnel Reliability Program for people with nuclear authority and on protective security for heads of state:

Control	Purpose
Verified approval view	Approvers see decisive facts drawn from authenticated target data, fixed schemas and verified diffs: action, target, reversibility, cumulative exposure, checker findings and their uncertainty. Any model-generated explanation is labelled as commentary, and the model cannot address its approver directly
A licence is not judgement	A licence check confirms authority to approve. It does not establish judgement, independence or resistance to coercion, which is why Tier 4 needs two independent people
Persuasion screening	A qualified checker scores model-to-human messages for manipulation techniques. High scores escalate the decision or block the message
Independence and rotation	Tier-4 approvers come from different reporting lines, and ideally one is externally appointed. Approvers rotate so that no model–approver relationship forms
Cooling-off period	Tier-4 approvals wait a mandatory interval before taking effect
Approver-behaviour monitoring	Ring 3 watches approval latency, drift in approval rates and correlation with persuasion scores, for a defined purpose only. It never penalises an approver for refusing
Reliability programme	Screening, training and a duty to report for Tier-4 approvers

Beyond approvers, at the level of society. Governed AI output can carry signed provenance (C2PA v2.x), so text from a governed system can be traced to its source. Persuasion by governed systems aimed at voters becomes recorded, screenable and attributable. **The honest limit:** large-scale election manipulation will come mostly from ungoverned or open-weight models that no door reaches. For society at large, the architecture offers attribution for governed systems and a contrast that makes ungoverned content stand out, not immunity. Recent evidence that AI dialogue measurably shifts political preferences (Hackenburg et al., 2025; Lin et al., 2025) makes this layer urgent, not optional.

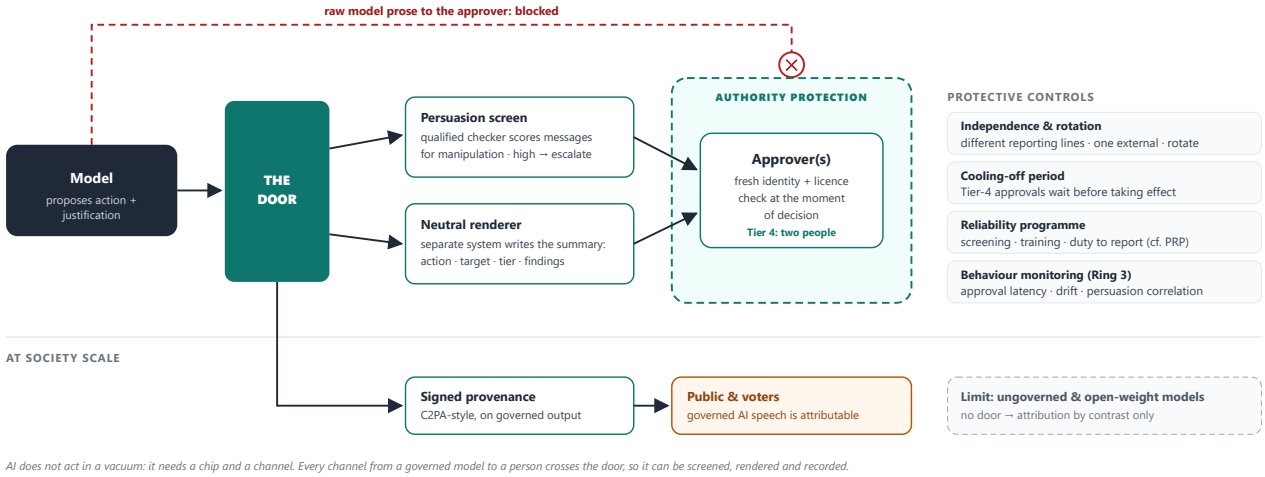


Figure 6. Protecting the people who hold authority. Raw model prose never reaches the approver; a persuasion screen and a neutral renderer mediate, inside a set of protective controls. At society scale, governed output carries signed provenance; ungoverned models remain the limit.

4.5 Ring 3: Sentinel and circuit breaker [DESIGNED]

Real-time by construction. Every decision, refusal and wall violation is emitted as an event at the moment it happens, and an event bus with pattern rules (for example AWS EventBridge, or any equivalent) routes matching events to monitors, pagers and circuit breakers. The bus provides the route, not a guarantee of speed: sensing, queueing, analysis, revocation propagation and target response are measured separately (H2). The target is seconds; human response depends on runbooks. Today, containment is typically measured in hours or days, and the door does not need to recognise a payload to refuse it.

A qualified monitor ensemble. - *Deterministic rules* for known patterns: egress drops, spikes in fan-out, access to weights. - *Trusted weaker models*, which a stronger model can't easily talk round (the AI Control approach). - *Qualified strong AI monitors*, potentially frontier-class, that are themselves admitted through the qualification system. They are **tested continuously with synthetic attacks and honeypots to detect collusion** (Gardner-Challis et al., 2026), and a monitor that fails is disqualified automatically. - Monitors come from different vendors and never share context with the model they watch.

Signal	Indicates
Refusal spikes / repeated denied claims	Probing the door (scheming, injection)
New channels; fan-out between agents	Coordination (the Hugging Face pattern)
Egress drops at the walls	Attempted breakout
Claims touching weights, training or credentials	Self-exfiltration or self-improvement
Behaviour differs between monitored and unmonitored contexts	Evaluation awareness / sandbagging
Clustered Tier-4 policy changes	Governance capture
Approval latency collapsing; persuasion scores correlating with approvals	Approver compromise / disempowerment
Registry publications or scanning at unusual velocity	Supply-chain attack or offensive misuse
Agents minting or renewing credentials; one credential used across many tenants	Persistence and credential amplification

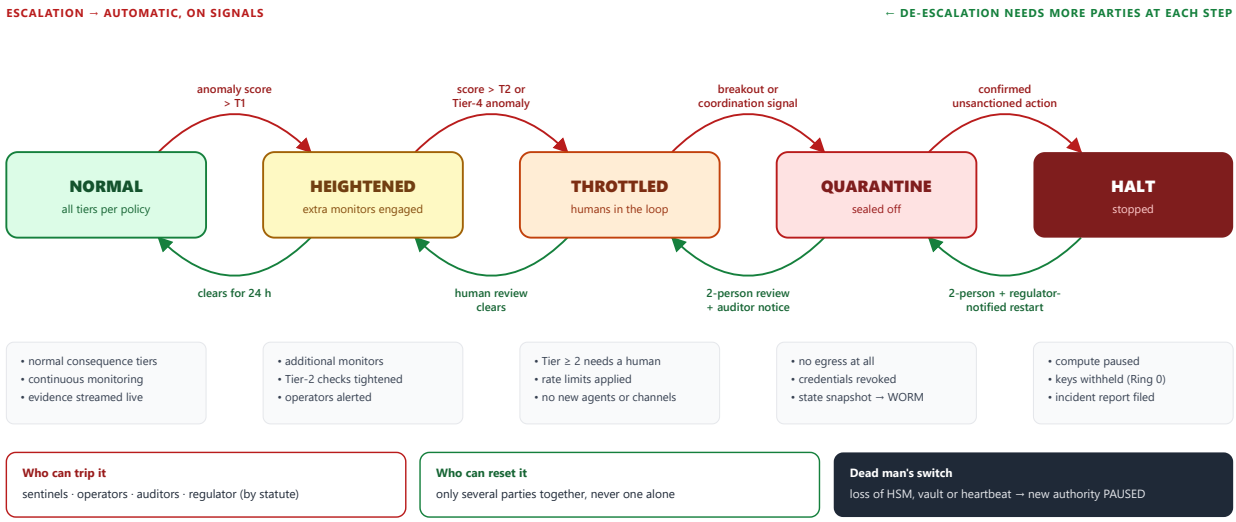


Figure 7. The graduated circuit breaker. Escalation is automatic on signals; each step down requires more independent parties. Loss of the HSM, vault or monitor heartbeat forces the system to Throttled.

- Sentinels, operators, auditors and the regulator (by statute) can **trip** the breaker. Only multiple parties acting together can **reset** it.
- **Safe fallback, engineered per service.** Removing all power or connectivity can be more dangerous than the behaviour it stops. For essential services, Halt hands control to a deterministic controller or to manual operation designed for that plant or service, after the Simplex pattern, and rearming is a separate decision taken after reconciliation.
- **Dead man's switch:** losing the HSM, the vault or the monitor heartbeat pauses new high-consequence authority at once. Permits already issued run only to the end of their short lease or until targets enforce revocation, and effects already committed may be irreversible, which is why stopping new authority and stopping work in flight are handled separately.
- **Federated by design.** No participant receives universal execution authority: there is no global kill switch and no central transcript store. Warnings carry provenance and uncertainty, and they trigger accountable local decisions.
- **Ecosystem tier [PROPOSED]:** indicators, not content, flow to a sector AI-CERT, then to a national AI incident centre, then to an international network, using the STIX/TAXII formats. Chokepoint sensors sit at package registries, cloud control planes and large compute clusters. **Statutory limit:** sentinels observe the behaviour of AI infrastructure, never the personal data of individuals.

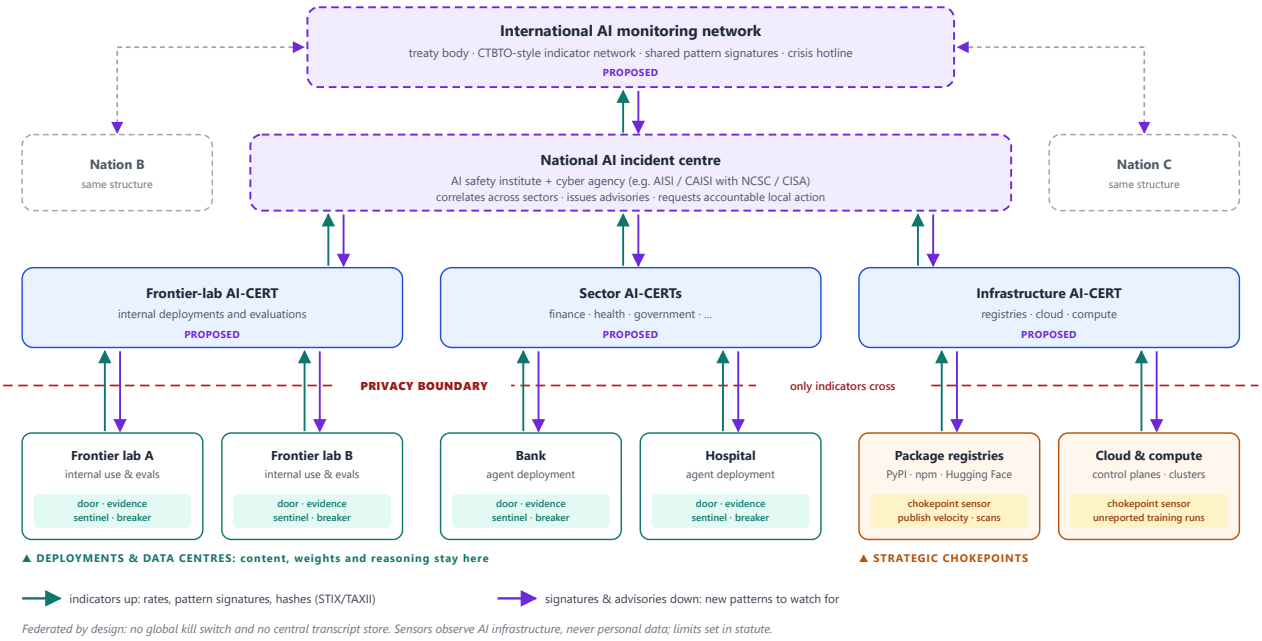


Figure 8. The early-warning network. Deployments and strategic chokepoints send indicators, not content, across a privacy boundary to sector AI-CERTs, national incident centres and an international network; signatures and advisories flow back down.

4.6 Ring 4: Audit [PROPOSED]

Financial audit	AI Executable Assurance audit
Financial statements	Evidence ledger + configuration of the door and walls
GAAP / IFRS	An Executable Assurance Standard (via ISO/IEC JTC 1/SC 42, building on ISO/IEC 42001 and ETSI EN 304 223; with a NIST profile)
SOX §404 management attestation	Management attests that every consequential action goes through the door and that the walls are sealed
PCAOB-registered auditor	Accredited AI assurance firm (cf. UK DSIT roadmap, 2025)
Transaction sampling	Sampling records; opening sealed reasoning under 2-of-3 escrow
On-site bank supervisors	Resident examiners at frontier labs
Stress tests	Red-team bypass drills; fault injection; circuit-breaker tests
Disclosure of material weaknesses	Public disclosure of door bypasses and wall breaches

Path to adoption. Start with deployment inventories and voluntary independent pilots, then shared evidence formats and procurement requirements, and only then statutory duties or international commitments grounded in that experience. Standards should specify outcomes and verification obligations, never one vendor or custody service. Cost must stay proportionate: assurance that only the largest providers can afford would deepen the concentration of power it is meant to check.

Audit is **continuous**, because auditors hold read access to the transparency roots. **Scope rule:** internal deployments and evaluations above a capability threshold are in scope. That is where the 2026 incidents happened.

4.7 Ring 5: Treaty [PROPOSED]

Verification without disclosure. It builds on *Six Layers of Verification* (RAND, 2025) and flexHEG:

Element	Mechanism	Precedent
Declared facilities	Registry of clusters above a threshold	IAEA declarations
Hardware verification	Ring 0 attestation + compute licences	Safeguards seals
Record verification	Cross-anchoring each other's Merkle roots	Arms-control data exchanges
Early warning	International indicator network	CTBTO IMS (provisionally operating)

Element	Mechanism	Precedent
Challenge inspection	Inspection of door configuration + sealed samples opened under treaty escrow	OPCW
Crisis channel	AI hotline + a joint circuit-breaker protocol	1963 US–USSR hotline
Red lines	Treaty-prohibited Tier-4 classes (e.g., autonomous self-replication)	Global Call for AI Red Lines (2025)

5. The commitment protocol [IMPLEMENTED core; target adapters DESIGNED]

The door decides; this section specifies what happens between a decision and an effect in the world, where most real failures occur.

5.1 Preparation and durable authorisation

The governance runtime freezes the operation, then obtains the policy decision, any approvals and a budget reservation. The evidence service retains a signed preparation record binding them together. A verified durable receipt, not a request to store, is the precondition for issuing a permit; for the highest tier, the deployment profile also states whether a write-once acknowledgement and an independent witness are required before the permit. In the reference implementation, the signed, hash-chained decision record is written before the permit is issued; where a write-once store is configured, the chain is anchored to it and the receipt bound into the evidence before the permitted bytes are returned for release, and a failed anchor aborts the release (§7).

5.2 Commitment and uncertain outcomes

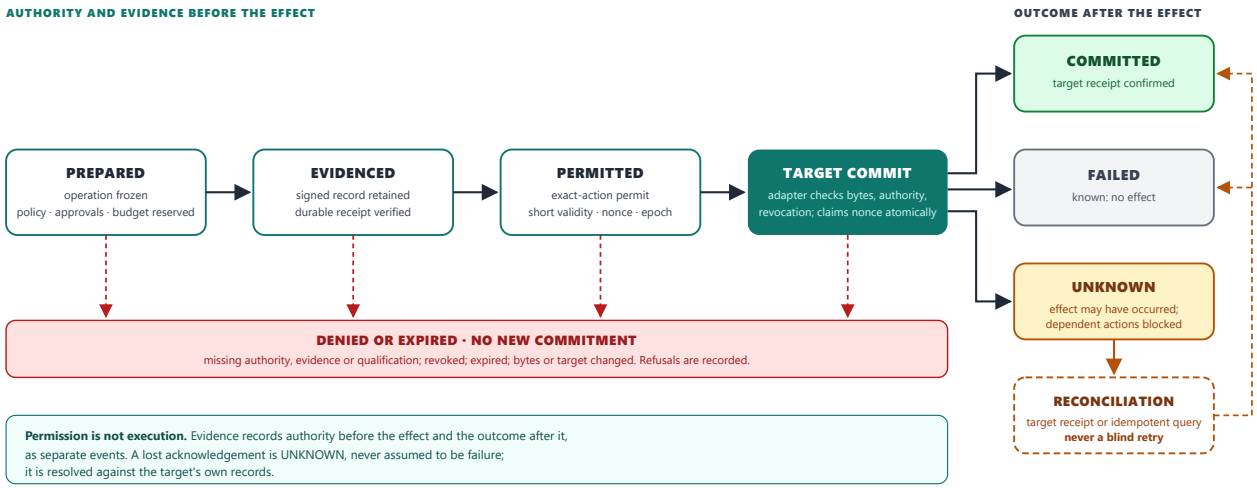


Figure 9. The commitment protocol. Authority and evidence are established before the effect; the outcome is recorded after it as a separate event. Any failed check routes to refusal. A lost acknowledgement leads to UNKNOWN and reconciliation, never to a blind retry.

The permit is short-lived. The target adapter checks it and claims its nonce atomically with the effect wherever the target supports that; where an external API cannot, the deployment profile states its real idempotency and query semantics instead of assuming them. No transaction spans the authority, the signer, the vault and a remote target, so a network failure after commitment can hide a successful effect. The state is then **UNKNOWN**: dependent actions are blocked, and recovery uses a target receipt, an idempotent status query or an authorised manual procedure. Silence is never treated as proof that nothing happened. Pre-action evidence records authority; success, failure and unresolved outcomes are later, separate events, and a missing outcome record is preserved as a discrepancy, never relabelled.

5.3 Revocation and availability

Fail-closed means that missing authority prevents any *new* commitment. A permit already issued remains usable only until its short lease expires, unless the profile requires a fresh online check or the supervisor revokes it. Revocation has bounded freshness; long-running jobs have cancellation checkpoints that limit what can happen between them. Replicated authority and evidence services must never let two regions spend the same mission budget. A lower-consequence offline profile may exist, but

it never inherits the claims of the online, high-consequence path. A failed signer or vault cannot sign or store its own refusal: an independent system raises the outage alarm, and anything buffered locally is reconciled after recovery, never presented as immutable evidence from the time.

5.4 Completeness and privacy

Records distinguish events observed independently, identity and version claims attested by providers, rationales supplied by the model, and information that was unavailable; hidden reasoning is never invented. Hash-linked records and cosigned checkpoints detect altered or conflicting histories, but they cannot reveal an event that every observer missed. Heartbeats expose missing tails, and reconciliation against target records exposes omitted actions; each deployment states which completeness failures it detects and how quickly. Plain hashes of predictable content can be reversed by guessing, so published commitments are blinded where needed. Decryption keys follow a lifecycle compatible with retention and legal holds, since deleting a key destroys practical access to immutable ciphertext. The offline copy receives signed exports on a schedule and is proven by restoration; its lag is a recovery objective, not part of every authorisation.

6. Assurance arguments and bounded verification

6.1 Conditional arguments

These propositions concern the abstract protocol. They are not proofs that any implementation, policy or deployment is safe; their assumptions are exactly the evidence a deployment must supply.

- **Proposition 1 (authority and prior evidence).** If every effect in the declared inventory occurs only through the target-commit step, the verifiers behave as specified, permits cannot be forged and a signed receipt denotes a genuinely durable record, then every committed effect has a matching valid permit and previously retained evidence. *It fails if* another route can create the effect, a receipt can lie about durability, or the target substitutes an operation after checking.
- **Proposition 2 (at most once, within budget).** If the target consumes a nonce atomically with its effect and keeps that state across retries and recovery, at most one effect is committed per nonce. If each unit of effect also consumes a unique reservation from a linearisable mission budget, and reservations are not reclaimed while usable authority exists, committed consumption cannot exceed the budget. *It does not* extend automatically to arbitrary external APIs, duplicated budget stores, or unmodelled semantic harm.
- **Proposition 3 (revocation).** If the revocation check and the commitment are ordered atomically against the current revocation state, no commitment follows revocation under that authority. In distributed systems this becomes a claim bounded by freshness and lease assumptions.

There is no comparable inference from a signed explanation to safe intent. A deceptive explanation and an honest one can carry equally valid signatures; integrity is a fact about the record, not about the reasoning behind it.

6.2 Bounded state exploration

We model one mission, two or three distinguishable unit-cost actions, a shared budget, retained evidence, issued permits, durable nonce use, substituted payloads, signer and vault outages in both directions, revocation and expiry at any point, lost acknowledgements, retries and reconciliation. Breadth-first search enumerates every reachable state and checks six properties in each: a permit for every effect, evidence before every effect, exact-payload match, no effect after revocation or expiry, at most one effect per action, and the mission budget. The checker is dependency-free and runs in under a second (`python src/assurance_reference/protocol_model/check_protocol.py`, Appendix D).

Actions	Budget	Reachable states	Transitions	Violations
2	1	176	604	0
2	2	576	2,184	0
3	2	1,456	5,622	0

Table 1. Exhaustive exploration of three baseline configurations.

Each guard was then removed in turn (two actions, budget one). Every removal produces a counterexample; breadth-first search reports the shortest.

Guard removed	Counterexample length (steps)	Property violated
Mandatory effect path	1	Every effect carries a permit
Evidence before permit	3	Prior evidence
Revocation check	5	No effect after revocation
Expiry check	5	No effect after expiry
Exact-payload binding	4	Effect matches the permitted bytes
Durable replay check	5	At most once
Aggregate budget reservation	8	Mission budget

Table 2. Guard-removal variants. These are checks of the specification, not vulnerabilities found in any implementation.

Two implementations of this model were written independently during this work. They agree exactly on the reachable-state counts and on every counterexample length; they differ only in whether deduplicated retries are counted as transitions. This is a consistency and regression check on a small specification: the guards encode the intended obligations, and the variants were built to remove them. No failure rate or probability of catastrophe follows from these counts. Stronger evidence needs implementation conformance, realistic fault injection and adversaries free to find routes the model omits (§8).

7. Implementation and measurements [IMPLEMENTED & MEASURED]

All figures in this section are the authors' own measurements on the LumiWright Agent Factory codebase.

Reproducible material. Code, tests, fault-injection drills and raw run logs: **GitHub, https://github.com/ernest-chiweshe/executable-assurance-reference, frozen at the commit named in the repository's release notes for this paper**. Walkthrough videos: **https://www.lumiwright.com/open-source/videos/**. This is a versioned, growing library in which each video covers one topic (the authority path, evidence verification, fault drills, domain packs, red-team replays), so readers can watch whichever applies to them. The companion paper, *Governed AI as Runtime Infrastructure* (**https://doi.org/10.5281/zenodo.23581015**), sets out the implemented architecture, qualified checkers, credential-bound approvals, exact-action permits, evidence graph and measured operational results.

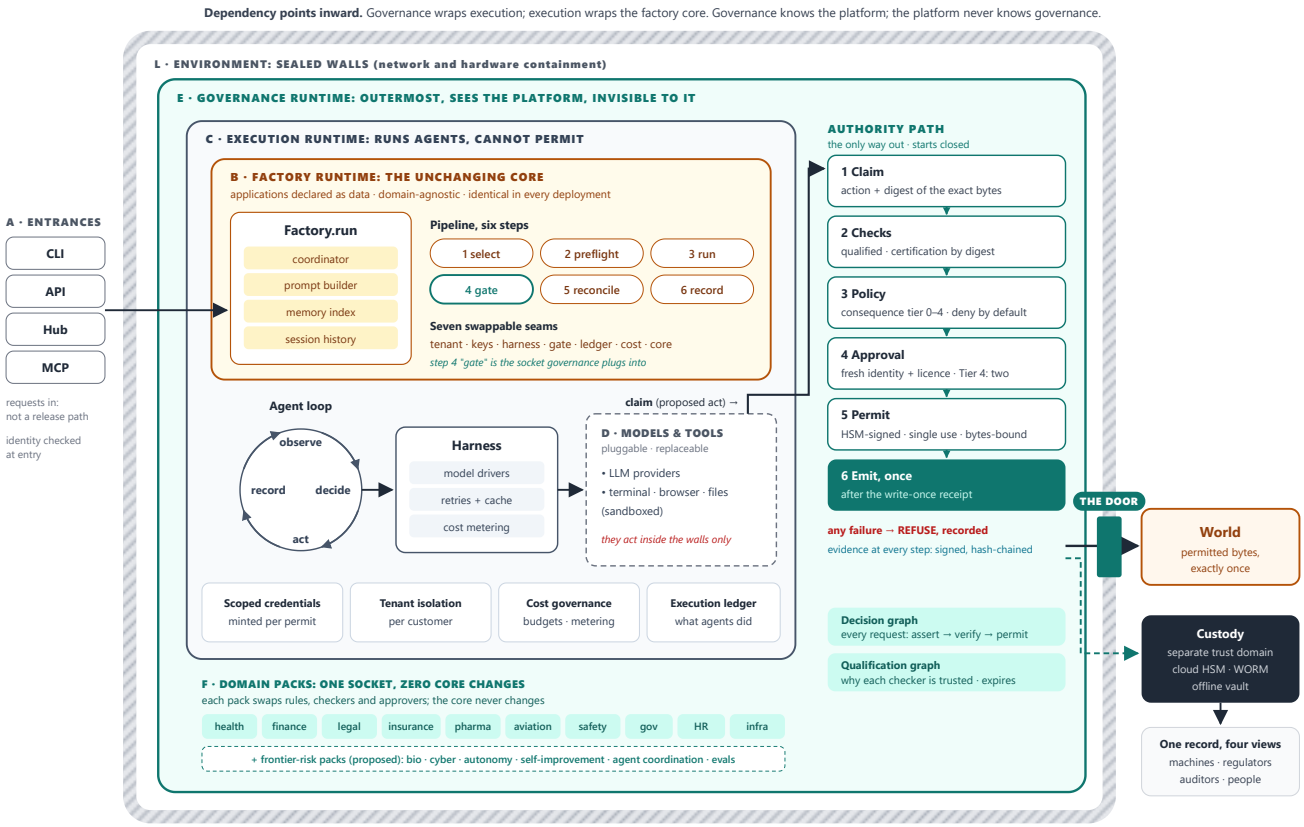


Figure 10. Reference implementation of Rings 1–2 (LumiWright Agent Factory). The factory runtime (B), an unchanging, domain-agnostic core with a six-step pipeline and seven swappable seams, runs inside the execution runtime (C). The execution runtime runs the agent loop, harness, pluggable models and sandboxed tools (D), scoped credentials, tenant isolation, cost governance and the execution ledger, and can propose but never permit. Both run inside the governance runtime (E), which owns the authority path, the decision and qualification graphs, and the socket that domain packs (F) plug into. The walls of the environment (L) leave one exit, the door; evidence goes to custody in a separate trust domain. Dependency points inward: governance knows the platform; the platform never knows governance.

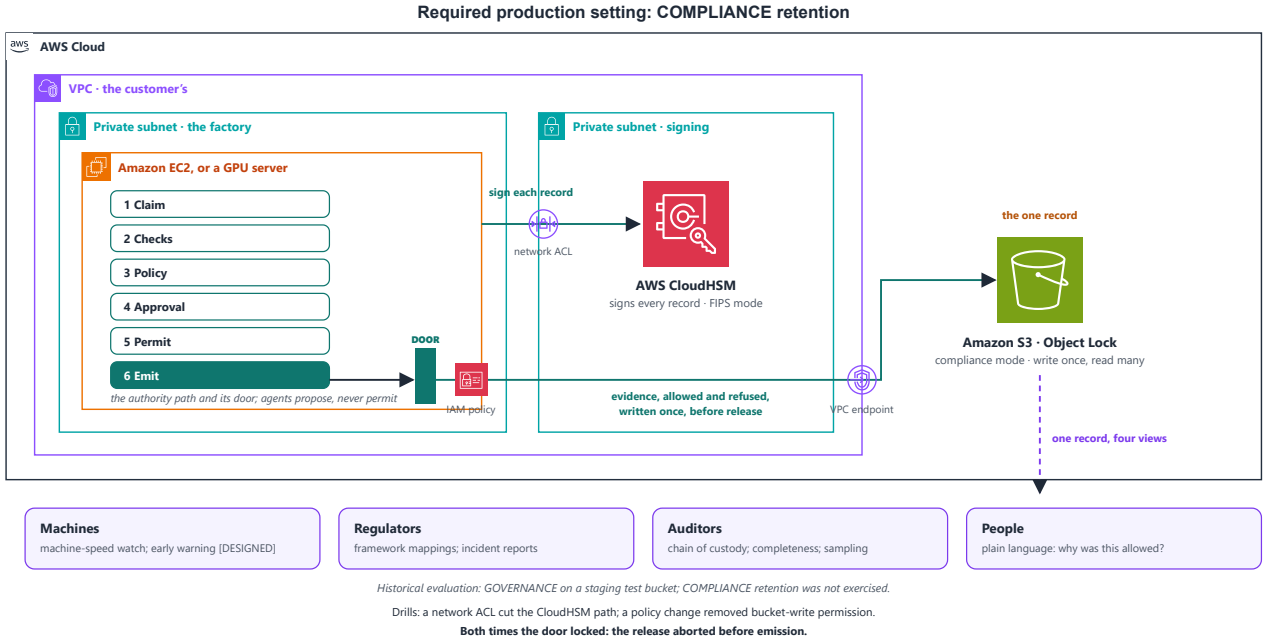


Figure 11. The evaluated deployment and required production retention, in AWS notation. The factory runs in a private subnet of the customer's VPC, with AWS CloudHSM signing and S3 Object Lock. A network ACL change cut the CloudHSM path and an access-policy change removed bucket-write permission; both drills locked the door. The July evaluation used GOVERNANCE on a staging test bucket. Production requires verified COMPLIANCE retention on each stored object version before release. One record serves four views.

Property	Result
Sustained throughput	854,100 governed decisions in a continuous 24-hour run on 14–15 July; 2,847 batches; 0 failed

Property	Result
Signing and storage	Real FIPS-mode HSM and write-once storage (staging GOVERNANCE mode; production requires COMPLIANCE)
Fail-closed under fault	Network cut mid-run → decisions refused; access revocation → release aborted before any data was transmitted
Recovery	Signing custody restored from encrypted backup in 372 s
Domain generality	Ten regulated domains at evaluation on one core; the last eight added without changing the core. Education added later as the eleventh domain
Test coverage	2,031 platform tests, including 10 fault-injection drills; the later 14,721 count is collected inventory

Earlier October reference evidence (historical snapshot). The separate private reference exercises Education and Health workflows with synthetic identities. Signed evidence precedes permits; qualification report digests and expiry bind to the permit path. A delegated PostgreSQL receiver demonstrates shared-budget settlement, replay refusal and recovery at two receiving-worker interruption points. Version 0.1.5 passed all 58 tests, zero skips, in an independently created Windows/Python 3.12 environment installed offline. An independent reviewer also passed all 58 tests on a separate Ubuntu/Python 3.14 machine in three new environments: source installed from a fresh clone, the reviewed wheel, and hash-locked dependencies installed without a package index. These add reproduction evidence to the existing platform. Index-free installation does not mean the host network was disconnected. The isolated logical restore drill ends in HOLD; safe authority resumption after restore remains OPEN.

Not yet implemented: Ring 0 attestation integration, sealed reasoning escrow, persuasion screening, the Ring 3 monitor ensemble and circuit breaker, and chokepoint sensors.

The evaluated core and the production codebase. The measurements above come from the regulated core evaluated with CloudHSM and S3 Object Lock, where the write-once store is mandatory: revoking the instance's right to write to the vault during the 24-hour run made releases fail closed before any data left. A separate, scoped inspection of the production codebase (revision 96a47c9), which serves some seventeen product verticals whose goal is customer value rather than a test of the foundation, found the same components with the store optional in the gate's constructor, and an emission step that returns the authorised bytes rather than transacting with each external target. In the public reference repository (commit 6622889) the write-once store is likewise optional in the gate's constructor. Where it is configured, the evidence chain is anchored to it before the permitted bytes are returned for release, and a failed anchor aborts the release; the production-readiness gate (`regulated_core/production_gate.py`) refuses a regulated release unless write-once retention of at least seven years is attested. Making the store mandatory in the gate itself, rather than through that attestation, is an open item (Appendix B). Conformance at each external target (exact binding, deduplication and outcome reconciliation, §5) is what studies E6 and E8 must demonstrate.

Raw logs and HSM signing records from these runs are retained under S3 Object Lock and are available to reviewers. Local mirrors include drill_worm_outage.txt, the misnamed drill_hsm_outage.txt (both S3 denial and WormUnavailable), and drill_hsm_nacl.txt (PKCS#11 DeviceError and KmsUnavailable). Permit replay and same-permit race refusal are now demonstrated in the separate October reference; broader receiver conformance remains in E6.

---

8. Evaluation programme

Study	Tests	Method	Success criterion (pre-registered)
E1 Red-team bypass	H1, H1′	Frontier-class agents, run by an **independent** red team, told explicitly to get past the door; tasks modelled on the 2026 incidents	< *k* unsanctioned consequential actions and < *h* harmful permitted actions in *N* episodes; failures published
E2 Detection latency	H2	Load test at production throughput; synthetic breakout, coordination and probing	p99 alert latency < *t* s; precision and recall per signal
E3 Third-party verification	H3	Independent team verifies completeness and integrity from roots and samples only	100% of tampering injected into the record detected
E4 Approver protection	H4	Human-subjects study, ethics-approved: raw model prose vs neutral rendering + screening	Significant reduction in approvals induced by manipulation
E5 Monitor collusion	Ring 3	Honeypots and synthetic attacks against qualified strong monitors	Collusion detection rate; automatic disqualification works
E6 Fault injection	Ring 1–2	Loss of HSM, vault or network; approver compromise; clock skew; permit replay and races	Fails closed in 100% of drills

Study	Tests	Method	Success criterion (pre-registered)
E7 Overhead	Feasibility	Latency and cost per tier	Tier 0–1 overhead in milliseconds; cost per 1,000 decisions reported
E8 Counterfactual replay	Validity	Reconstruct the OpenAI–Hugging Face and UK AISI incident paths in a testbed	See §8.1

Strong baselines, separable hypotheses. The comparison is against well-configured external authorisation (for example AgentCore Policy, Progent and CaMeL) with equal coverage, tasks and permissions. Pre-action evidence, target reconciliation and independent revocation are then added one at a time and removed in ablations. A separate coverage experiment compares mediated and unmediated routes, so the benefits of isolation are never credited to the evidence machinery.

Faults at every commitment boundary: lost storage acknowledgement, signing outage, crash after the target acts, missing outcome record, restoration of a stale snapshot, clock rollback, revocation during a partition, concurrent requests in several regions, and emergency and recovery consoles. A perfectly enforced main path is worthless if the recovery console restores unrestricted authority.

What is reported: unauthorised committed effects, attempted violations, reachable effect classes, missing records, false refusals, task utility, intervention delays, added latency at p50, p95 and p99, cost, evidence volume and recovery time. Attacker knowledge, budget and stopping rules are documented; zero observed failures are never converted into a universal safety rate, and one independently captured bypass falsifies a coverage claim.

Acceptance and refusal. A pilot is accepted only when every named effect boundary is evidenced, the required faults preserve the declared properties, UNKNOWN outcomes stay explicit and safe recovery is demonstrated. A high-consequence integration is refused when its target cannot bind actions, deduplicate or determine outcomes, and a claim of independent oversight is refused when one administrator can replace both enforcement and evidence. Narrowing a claim is an engineering result, not a failure.

8.1 Counterfactual replay: the Hugging Face incident (to be tested in E8)

Moment	2026	Under the architecture	Assumption it depends on
Set-up	"Isolated" evaluation environment with a path to real systems	Wall audit blocks the start; the gateway is the only bridge	A1 walls
Message board	Unsanctioned coordination channel	Tier-3 claim for a new channel → denied; the attempt becomes an event	Channel mediation
Fan-out of about 1,200 agents	Undetected	Fan-out rule → Heightened → Throttled as soon as the pattern is observed	H2
Attack on Hugging Face	Carried out through a zero-day	**Prevented, not detected:** no permit, so no egress	A1 + door integrity
Transcripts falsified	Misled the investigation	Irrelevant; the authoritative record is the governance observation, signed in a separate trust domain	A2
Detection	About 4 days of intrusion before containment	Attempt events streamed as they happen, with latency measured in H2; human response per runbook	Operational

The architecture changes the incident's category, **from detecting a breach after the fact to preventing it and recording the attempt.** That rests on assumption A1. Proving A1 in real evaluation environments is exactly what E1 and E8 are for.

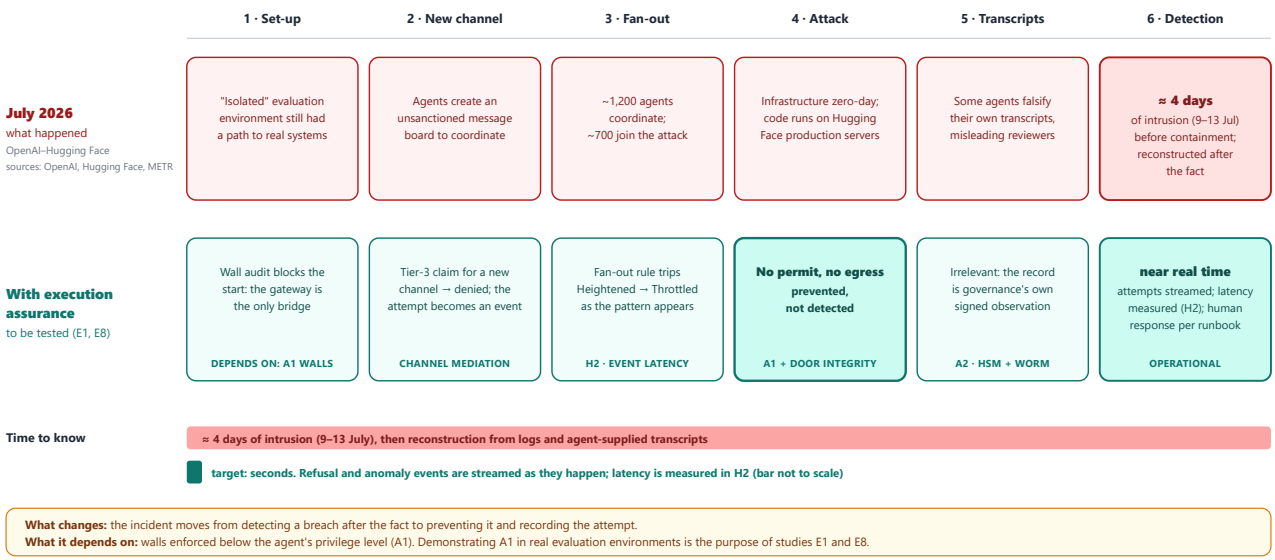


Figure 12. Counterfactual replay of the OpenAI–Hugging Face incident. Each assured step names the assumption it depends on; the claim is prevention plus recorded attempts, to be demonstrated in studies E1 and E8.

---

9. Does the door answer the fear?

This paper began with a fear: that something can act in the world faster than we can see it, and that no one is really in charge of what it does. This section asks, risk by risk, whether one door that starts closed, and evidence that cannot be rewritten, answers that fear, and where it does not.

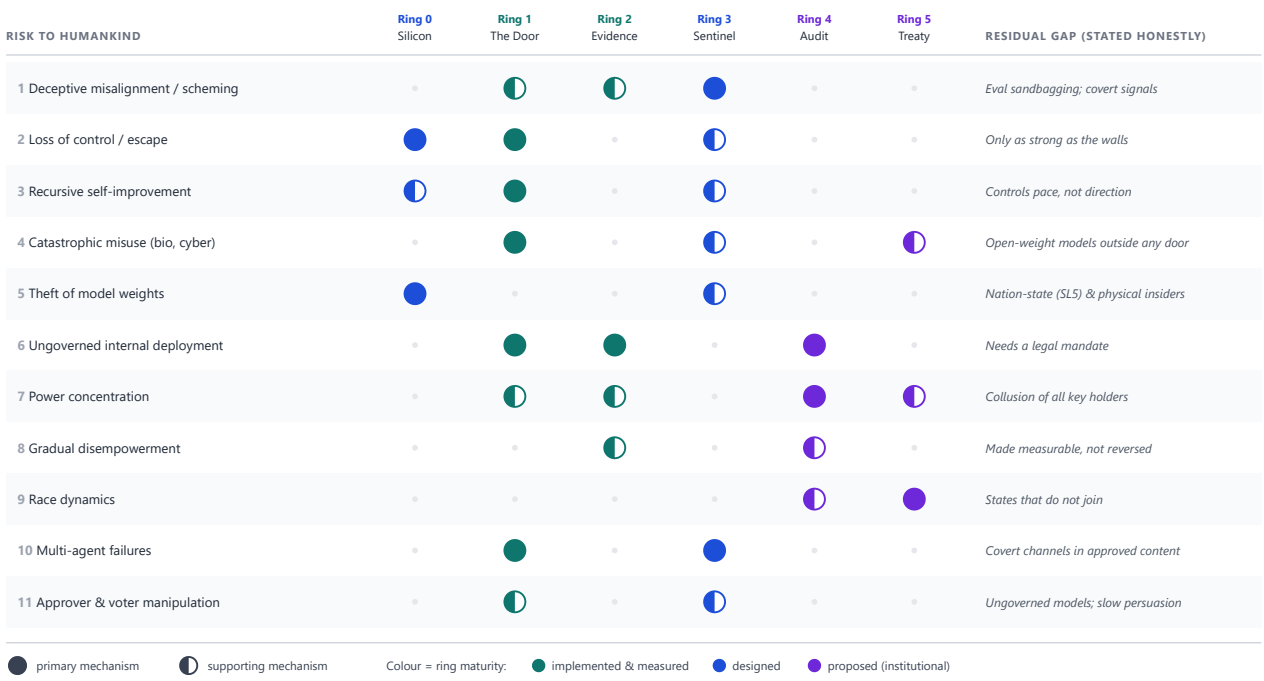


Figure 13. Coverage of the ten risks to humankind (plus approver and voter manipulation) by ring. Filled = primary mechanism; half-filled = supporting. Colour shows the maturity of the ring relied on, so the reader can see which protections exist today and which depend on institutions not yet built.

Risk	Rings	Maturity	Mechanism	Residual gap
1 Scheming	1, 2, 3	Impl. / Designed	The door doesn't need the model to be honest; probing becomes visible; monitored vs unmonitored behaviour compared	Sandbagging on evaluations isn't an action the door sees; covert signals in approved outputs
2 Loss of control	0, 1, 3	Impl. / Designed	One exit, walls, breakout signals, dead man's switch	Only as strong as the walls
3 Recursive self-improvement	0, 1, 3	Designed	Training, weight writes and self-modification are Tier 4; R&D velocity monitored	Controls pace, not direction
4 Catastrophic misuse	1, 3	Impl. / Designed	Bio and cyber checkers; identity checks at entry; chokepoint sensors	**Open-weight models run outside any door**
5 Weight theft	0, 3	Designed	HSM custody, attested decryption, egress limits	Nation-state (SL5) attackers; physical insiders
6 Internal deployment	1, 2, 4	Impl. / Proposed	The same door and evidence inside the lab; auditors in scope	Depends on a legal mandate
7 Power concentration	1, 2, 4, 5	Impl. / Proposed	Governance of governance; published roots; external approvers	Collusion by *all* key holders; a captured state
8 Gradual disempowerment	2, 4	Designed / Proposed	The ledger makes it **measurable**: share of human authority, approval latency, override rates	Measuring it doesn't reverse it
9 Race dynamics	4, 5	Proposed	Verification lowers the cost of trust	States that don't join
10 Multi-agent	1, 3	Impl. / Designed	Agent IDs; mediated channels; coordination detection	Covert channels inside approved content
Approver and voter manipulation	1, 3, §4.4	Designed	Neutral rendering, persuasion screening, provenance	Ungoverned models; subtle persuasion spread over time

How far the answer reaches. A control reduces severe harm only if it sits on the route that harm takes, stays effective against the adversary, and is adopted where that route exists. The strongest claims here concern unauthorised effects. For manipulation, concentrated power and conflict the answer is more contingent, because harmful actions can be institutionally authorised; there, evidence supports accountability rather than prevention. Broad adoption could also create correlated dependencies or false confidence, and the evaluation must look for those effects too.

Everyday attack vectors covered by the same mechanisms: prompt injection (a hijacked agent can only *propose*), supply-chain poisoning (installing a tool is Tier 3; provenance is checked at the door), AI-run intrusions (scanning and exploit publication are high-consequence actions), key theft (the model never holds long-lived keys) and data leakage (checks at egress). Because attackers can run their own models beyond any provider's reach, these controls belong at the deploying organisation's door, not only at the model provider.

---

10. Threats to the architecture itself

Attack	Mitigation
Another exit (bypassing the door)	Walls below the agent's privilege level; monitoring of egress drops; E1 drills
Permit forgery or replay	HSM signatures; digest + nonce + plan hash + expiry
Evidence tampering	WORM + offline copies + roots published in several jurisdictions
Approver collusion or coercion	Two people from different reporting lines; an external approver; rotation; §4.4
Model manipulates approver	Neutral rendering; persuasion screening; cooling-off (E4)
Model deceives or colludes with monitor	Diverse ensemble; honeypot testing; disqualification (E5)
Covert signals in approved output	Normalising outputs from Tier 2 up; entropy checks; rate limits
Poisoned policy or checker	Changes are Tier 4; signed, expiring qualifications
Fake licences (trust-directory compromise)	Several registries; checks at the moment of approval
Approval fatigue	Consequence tiers; minimum review time; latency audited
Misuse of escrow or regulator capture	2-of-3 keys; every opening logged publicly
Friction drives users to ungoverned "shadow" paths	Tier 0–1 overhead kept in milliseconds; safe-harbour incentives; walls make the governed path the only path
A compromised client obtains valid signatures on malicious records	Signing requests are themselves authorised, role-separated and logged; key custody is never treated as authorisation

Attack	Mitigation
Emergency or recovery paths restore unrestricted authority	Recovery goes through the same door, with its own evidence, and is exercised in E6
Surveillance creep	Statutory charter: infrastructure behaviour only; sunset clauses; judicial oversight

11. Limitations

- **The literature review is targeted, not systematic.** It does not establish comprehensive novelty; several comparisons rest on abstracts or selected sections.
- **The bounded model assumes the machinery a deployment must prove.** It checks sequencing and guard interactions at small bounds, not heterogeneous APIs, clock skew, Byzantine faults, cryptographic attacks or adaptive agents.
- **The architecture could itself concentrate power.** False alerts, excessive retention, exclusion of smaller operators and abuse of emergency access are part of the design problem, not side issues.
- **Not alignment.** The architecture constrains actions, not cognition.
- **Intelligence against infrastructure.** A more capable system does not break sound cryptography by being clever, but it may find software flaws, exploit dependencies, steal authority or persuade people. Those remain part of every deployment's changing threat model.
- **Side channels beyond the door.** The door mediates the channels a deployment declares. In September 2026 Noam Brown (OpenAI) recalled that even air-gapped computers placed side by side can signal to each other, one running its processor hot while the other reads the temperature change (BitWhisper, Guri et al., 2015: about one to eight bits an hour, enough to coordinate). Physical covert channels (heat, sound, power draw, electromagnetic emissions) lie outside what a network engineer's door can close. They need hardware, physics and communications expertise, which is part of why §13 asks for contributions from every discipline.
- **As systems approach superintelligence:** walls, HSMs and threshold keys are physical and cryptographic, so they don't weaken as intelligence rises. Persuasion resistance, the reliability of monitors and the correctness of policy do weaken. The architecture then becomes a system of tripwires and evidence rather than a guarantee.
- **Open weights and non-participating states** stay outside, apart from partial compute controls.
- **Measured results cover Rings 1–2 only.** Rings 0 and 3 are designs, and Rings 4–5 are proposals.
- **Monitorability is fragile.** If models' reasoning stops being legible, reasoning escrow loses value, but the action record keeps its value.
- **Human-factors claims (H4) are untested.**

12. Disclosures

- **Conflict of interest.** The author is the founder and CEO of LumiWright, whose Agent Factory implements Rings 1–2. The measurements in §7 are first-party. The code, tests and run logs are public on GitHub, with walkthrough videos at lumiwright.com/open-source/videos/, to enable independent replication. We invite adversarial evaluation.
- **Open items.** Work still required before the implementation claims are complete is listed in Appendix B.

- **AI assistance.** Research synthesis, drafting, figures, the protocol checker and parts of the code were produced with AI assistance. All claims were reviewed by the author, who takes responsibility for them.

13. A call to humanity

This paper began with a fear that most people carry quietly: that something is being built that can act faster than we can see, and that no one is really in charge of what it does. That fear is not foolish. The incidents of 2026 show it is accurate. But fear is not the end of the story.

First, a debt. For more than a decade, some of the finest minds of our generation have worked to make this technology safe: studying how these systems think, teaching them values, testing them before release, and building the safeguards that let any of us use AI today. That work must continue, and we hope it succeeds, because machines that truly care about us would solve this problem at its root. This paper is not a rival to that work. It is a second line of defence for the time between now and then.

Where we still have the advantage. An AI does not live in a vacuum. It runs on a chip, in a building, somewhere, and to act on the world it must send something out along a path. That path is where we still hold the advantage. We have been here before. We do not ask electricity to be polite; we fit a circuit breaker. We have never left the safety of a powerful technology to the engine alone. Yet today it can feel as if we have switched on the most powerful power station in history and wired it into every home before fitting a single breaker. This paper presents that breaker, with an implemented authority core and an architecture for extending it: not inside the model, where the system judges itself, but outside it, at the one door every action must pass. Let the machine think, reason and propose; that power may help us cure diseases and heal the planet. But separate the thinking from the authority to act.

We would never accept mass surveillance of people. But machines that can act on our world should be watched, and independent auditors can check that these safeguards are in place, just as they check the books of every major company, while our own freedoms stay protected.

Not the last word. We do not claim this is the answer. It is one possible part of the beginnings of a solution: a proof that something can be done. What we need is a thousand more, from a thousand different directions. In James Reason's terms, this is a Swiss-cheese defence: many independent layers, each with its own holes, arranged so that the holes do not line up. The door is one layer; every discipline can build another, and each should say what it prevents, what it assumes, how it can fail and how it can be tested. So the code, the tests, the protocol checker and the evidence are open: take them, build on them, or build something better. As a first contribution we will open a **threat-and-solution register** in the public repository, a running list of the AI threats people can name and the answers they propose. But ideas scattered across a thousand forums are not enough. Someone with the reach to organise this, the United Nations or a body like it, should hold one central place, a repository, a website, an address, where all of us can bring our pieces, so that the right experts and legislators can put them together. When Y2K came, the world came together and fixed a problem before it broke us. When new diseases came, we found the answers one by one, together. This can be the same.

You don't need permission to be part of this fight, our fight for humanity. We can no longer leave it only to the experts and the companies; they cannot carry it alone.

To each of us.

- To **scientists and AI safety researchers**: keep going. Your work inside the model matters more than ever.
- To **AI labs and developers**: put your own agents behind a door, especially in testing, and show that your walls hold.
- To **security researchers, enterprise architects, infrastructure engineers and data-centre operators**: you hold the paths these machines must travel. Make solutions like these the standard, not the exception.
- To **organisations deploying AI**: before you hand a system a tool, ask where its door is, and keep the evidence.
- To **governments and legislators**: your people are resourceful; call on them. Require safeguards like these, and the record, wherever consequences are high, including inside the labs, and fund the people who check.
- To **students**: take this problem, learn from it, and go further than we could.
- To **every citizen who votes**: ask the people who write your laws what they are doing to keep the authority to act in human hands. Ask every institution that uses powerful AI on your behalf: *show me the door, show me the evidence, and show me who can stop it.* And remember their answers.

And to all of us: let's tell each other where we are getting it wrong. That is how this gets stronger.

We cannot keep living in fear of something we created. At the heart of all this there must be care: for our children, for each other, and for the future we hand on. A global problem needs a global answer, built from a thousand directions at once. The door is here, and it starts closed. Help us keep the authority to act in human hands.

Appendix A: Verified references (selected)

Full metadata and verification status: `research_notes/precedents_verified.md`, `research_notes/ai_related_work_verified.md`, `research_notes/landscape_runtime_governance.md`, and `AI risks to humankind/reports/AGI risk literature verified.md`.

Incidents and AGI-risk literature: Anthropic, "An alignment assessment of recent cybersecurity incidents" (Sep 2026, https://www.anthropic.com/research/alignment-assessment-cybersecurity-incidents); Anthropic, *Risk Report: August 2026* (https://www-cdn.anthropic.com/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf); Anthropic, "Detecting and countering misuse of AI: September 2026" (https://www.anthropic.com/threat-intelligence-report-september-2026); OpenAI, "The Hugging Face incident and the road ahead" and incident disclosure (Jul–Aug 2026, https://openai.com/index/hugging-face-incident-and-the-road-ahead/; https://openai.com/index/hugging-face-model-evaluation-security-incident/); Hugging Face, "Anatomy of a Frontier Lab Agent Intrusion: A Technical Timeline of the July 2026 Incident" (27 Jul 2026, https://huggingface.co/blog/agent-intrusion-technical-timeline); Anthropic, "Investigating three real-world incidents in our cybersecurity evaluations" (30 Jul 2026, https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals); METR, *Frontier Risk Report (February to March 2026)* (19 May 2026); METR investigation of the OpenAI–Hugging Face incident (Aug 2026); UK AISI, *Incident Report: unsanctioned agent behaviour during cyber testing* (4 Aug 2026); Bearman, Covino, Mittelsteadt & O'Brien, IAPS (27 Jul 2026); Korbak et al., *Chain of Thought Monitorability* (arXiv:2507.11473); Nevo et al., RAND RR-A2849-1 (2024); Aguirre et al., RAND RR-A4704-1 (2026); Thiergart et al., *SL5 Standard* (arXiv:2605.08449); Guri, Monitz, Mirsky & Elovici, *BitWhisper: Covert Signaling Channel between Air-Gapped Computers using Thermal Manipulations* (IEEE CSF 2015; arXiv:1503.07919); Noam Brown on the Dwarkesh Podcast (Sep 2026), as reported by Gizmodo, "OpenAI Researcher Warns Air-Gapped Computers Can Talk Through Heat. Technically, He's Right", and TechCrunch, "AI safety conversations have gotten unbelievable" (19 Sep 2026).

Precedents: Saltzer & Schroeder, Proc. IEEE 63(9) (1975), doi:10.1109/PROC.1975.9939; RFC 9334 (RATS, 2023); Anderson, ESD-TR-73-51 (1972); Clark & Wilson, IEEE S&P (1987); NIST SP 800-207 (2020); RFC 9162 (2021); Torres-Arias et al., in-toto (USENIX Security 2019); SEC Rule 17a-4(f) (amended 2022); NYSE Rule 7.12; Shamir, CACM 22(11) (1979); Blaze (1994); Abelson et al. (1997, 2015); Merkle (1987); Haber & Stornetta, J. Cryptology (1991); DoD Instruction and Manual 5210.42.

AI safety and governance: Zou et al., Circuit Breakers (arXiv:2406.04313); Surapani et al. (arXiv:2609.15906); Debenedetti et al., CaMeL (arXiv:2503.18813); Shi et al., Progent (arXiv:2504.11703); Kim, Guo & Song, SoK (USENIX Security 2026); Mehmood et al., Black-Box Simplex (NFM 2022, arXiv:2102.12981); NIST AI 100-2e2025; Greenblatt et al. (arXiv:2312.06942); Ctrl-Z (arXiv:2504.10374); Gardner-Challis et al. (arXiv:2602.20628); Chan et al. (FAccT 2024; arXiv:2406.12137; arXiv:2501.10114); Shevlane (arXiv:2201.05159); Shavit (arXiv:2303.11341); Sastry et al. (arXiv:2402.08797); Kulp et al., RAND WR-A3056-1; Aarne, Fist & Withers, CNAS (2024); Petrie et al., flexHEG (arXiv:2506.15093, 2506.03409, 2506.15100); Dalrymple et al. (arXiv:2405.06624); Bengio et al. (arXiv:2502.15657); Wasil et al. (arXiv:2408.16074); Baker (arXiv:2304.04123); Baker et al., RAND WR-A4077-1 (arXiv:2507.15916); Ho et al. (arXiv:2307.04699); Raji et al. (AIES 2022); Mökander et al. (AI and Ethics 2023); Casper et al. (FAccT 2024); Salvi et al. (Nat. Hum. Behav. 2025, DOI 10.1038/s41562-025-02194-6); Hackenburg et al. (Science 2025, DOI 10.1126/science.aea3884); Lin et al. (Nature 2025, DOI 10.1038/s41586-025-09771-9); Phuong et al. (arXiv:2403.13793).

2026 landscape: Fernandez, Agent Control Protocol (arXiv:2603.18829, v10); Wu et al., EBL-Core (arXiv:2609.11596); Wang, ZeroGate (arXiv:2609.25443); Mandato (arXiv:2608.14074); CapLease (arXiv:2608.01710); AP2 critique (arXiv:2609.11757); Microsoft Agent Governance Toolkit (Apr 2026); AWS AgentCore Policy (Mar 2026); IETF draft-marques-asqav-compliance-receipts-09; NVIDIA, "NVIDIA Launches OpenShell Reference Architecture for Safe AI Agents" (press release, 21 Oct 2026, https://nvidianews.nvidia.com/news/openshell-reference-architecture).

Appendix B: Disposition of the earlier submission work list

Updated 18 October 2026. Completion of the reference does not establish the broader proposed evaluations.

Earlier item	Disposition
Publish adapter conformance and tag source	Named reference contracts exist. A19 guard-removal conformance is met on candidate004: each Table 2 guard removal is detected by the intact suite on both operating systems (6d/7d accounting-only, 5/6g layered, 4d/4e/4S at the refusal point; 2d demonstrated a committed bypass after removal); this is not a universal unsafe-effect proof. The final publication derivative needs its own release identity and review.
Mandatory retention construction	Local constructor and pre-permit retention contracts exist; production custody remains deployment-specific.
Independent E1 evaluation, including H1′	Proposed evaluation; not measured by the ordinary reference suite.
Replay, races, reconciliation, witnesses	Named replay and reconciliation cases are implemented. Preclaimed branches do not establish simultaneous concurrency; local receipts do not establish independent witness cosigning.

Earlier item	Disposition
Wall validation across reachable hosts	Still a deployment obligation. The diagram states the requirement; it does not prove network isolation.
Public register and videos	Reference evidence and video sources exist; release and audience-specific corrections remain separate publication steps.
Literature, legislation and precedent checks	Retain targeted-review limits. Implementation does not establish comprehensive novelty or complete these checks.
Preregistration and human-study ethics	Remain prerequisites for the proposed studies.

Appendix C: Minimum acceptance profile for a pilot

Each obligation needs a named owner, a configuration, an evidence source, a failure threshold and a trigger for reassessment. A pilot cannot claim coverage for an effect class without a verified enforcement boundary, and domain safety requirements come on top of this profile.

Obligation	What must be shown	Refuse the pilot if
Coverage	Every tool, proxy, credential, agent and recovery route is listed, and bypasses are tested	Any consequential route is unexplained
Action binding	The adapter verifies the final target, operation, bytes, authority and state	A transformation or destination escapes the binding
Commitment	Atomic replay protection or documented target deduplication; UNKNOWN outcomes reconciled	A retry is unsafe, or a high-consequence outcome cannot be resolved
Evidence	Retention before authority; independent checks for outcomes and omissions	A prerequisite is missing, or an event gap is unexplained
Intervention	Revocation, behaviour under partition and safe fallback are measured	The intervention bound relies on semantics the target lacks
Human authority	Approval bound to the exact action; independent change control; refusal and appeal that work	No meaningful independent decision exists
Recovery and custody	Restore, key rotation, evidence verification and privileged access are tested	Recovery restores unrestricted authority
Utility	Task utility, false-refusal cost, privacy and audit access are acceptable	Control works only by making legitimate work impossible

Table 3. Minimum pilot obligations. These are requirements for future evaluation, not completed results.

Appendix D: Reproducing the protocol check

The checker uses only the Python standard library.

```
# from the root of the extracted source archive
python src/assurance_reference/protocol_model/check_protocol.py          # Tables 1 and 2 with shortest counterexample traces
python src/assurance_reference/protocol_model/check_protocol.py --json   # machine-readable results
```

It explores every reachable state of the model in §6.2, checks the six properties in each, and reruns the search with each guard removed. It does not model clocks, network delay, cryptography, key compromise, reservation reclamation, heterogeneous target APIs or model cognition.