ERES-ARH-2026-001 — Adversarial Root Hardening

Role: Root Threat-Model · Independence Gate · Adversarial Harness **Version:** v0.1 PROPOSED — pre-canonical, NOT MIEVM-rated **Author:** Joseph Allen Sprute ("ERES Maestro") · ORCID 0000-0001-9946-3221 **Institute:** ERES Institute for New Age Cybernetics · Bella Vista, Arkansas **License:** CC BY 4.0 / CCAL v2.1 **ID status:** pending JAS confirmation

Abstract

The ERES safe root install specifies a seating order — an immutable Floor (the three constitutional rules under Floor Supremacy), an append-only provenance spine, a Human-to-Computer-to-Human topology that denies the machine an endpoint, a conjunctive admissibility gate, a vocabulary firewall, a ratification quorum, and non-punitive remediation — installed below the application so that containment is topological rather than reactive. That specification is internally coherent, but the stronger claim "prevents malicious activity" is not yet earned: it names no adversary and demonstrates no layer actually refusing an attack. This instrument closes that gap to an explicit boundary, and contributes three artifacts.

First, a threat model that declares exactly one attack per root layer (T1 privilege-escalation on the Floor; T2 forged or truncated provenance; T3 spoofing or coercing the human terminus; T4 admissibility-gate evasion; T5 redefinition of a load-bearing token; T6 node collusion/sybil; T7 striking during the remediation window) and, for each, the mechanism that refuses it. Named attacks are the adversarial equivalent of the falsifiers the Doctrine of Verifiable Claims requires: a hardening that cannot state its attacks has not earned the word "prevents."

Second, an Independence Gate that closes T6 by reusing an existing ERES instrument. It wires the Covariance-Discounted Floor Engine's effective- independence count, n_eff — computed as (sum of sqrt of eigenvalues)² over their sum, taken over the node verdict-correlation matrix — into ratification as an admissibility precondition conjunctive with the MIEVM convergent-mean gate. Because correlated nodes collapse n_eff toward one regardless of how many are run, a manufactured "convergent mean" is refused by construction; the sybil route is closed structurally rather than detected after the fact. The gate inherits CDFE's Assumption A1: n_eff on verdict covariance measures apparent, not true, independence.

Third, a dependency-free adversarial harness (including a pure-Python eigensolver, so the gate is computed rather than stubbed) that throws each attack at a minimal root model and passes only when it is blocked; five enumerated OPEN items remain skipped until closed. A green run is evidence about the specification against a modeled adversary — not about a live runtime against a novel adversary, which stays gated on external red-teaming (item O5, which simulation cannot discharge). The instrument tests no claim about its own goodness and exposes no self-certification verdict.

0. Purpose

The ERES safe root install specifies a *seating order* — Floor, provenance, topology, admissibility, vocabulary firewall, ratification, remediation — that makes containment topological rather than after-the-fact. That specification is coherent. What it lacks, to support the stronger claim "prevents malicious activity online and anywhere else ERES operates," is a **named adversary** and a **demonstration that each layer refuses its attack.**

This instrument closes that loop to a stated boundary. It supplies three things:

1. a **threat model** — one declared attack per root layer (T1–T7);
2. an **Independence Gate** — the ERES-CDFE `n_eff` scorer wired into ratification as an admissibility precondition, closing the collusion/sybil route (T6);
3. an **adversarial harness** — a stdlib-only executable that throws each attack at a minimal root model and goes green only when every modeled attack is blocked and every OPEN item is closed.

It does **not** claim to have secured a live runtime. Section 6 fixes that boundary explicitly.

---

1. The root, restated as an attack surface

The install seats seven layers, each depending on the one beneath it. Read as an attack surface, each layer is defined by the one thing an adversary would do to defeat it.

#	Layer (root instrument)	Function	Declared attack
1	DICTUM / empirical Floor (three rules; Floor Supremacy)	immutable base constraint	**T1** privilege escalation to edit the invariant
2	Provenance spine (SOMT append-only log)	origin-anchor every act	**T2** forged provenance / append-only bypass
3	H2C2H topology (RCR + RT qualifier)	deny the machine an endpoint	**T3** spoof or coerce the human terminus
4	PoR admissibility (A×R×P×F)	block inadmissible actions	**T4** gate evasion — non-zero on all factors yet malicious
5	Floor Register (vocabulary firewall)	pin load-bearing meaning	**T5** redefinition — drift a token so the old rule "passes"
6	Ratification (MIEVM + CBGMODD)	no unilateral root change	**T6** node collusion / sybil — manufacture a convergent mean
7	Remediation (Paineology L3→L1)	contain, don't destroy	**T7** strike during the downgrade window

Figure 1 — The seven-layer ERES root as an attack surface, with the Independence Gate (T6) and the claim boundary.

The DVC's element 1 (falsifiability) requires named falsifiers. T1–T7 are the adversarial equivalent: named ways the safety claim could be shown false. A hardening that cannot state its attacks has not earned the word "prevents."

2. Blocking mechanism, per layer

- **T1 — Floor edit.** Floor Supremacy is a constitutional invariant *about* Row-1 quantities; it is not an application setting. In the model the Floor is locked below the mutable surface: an edit call is refused regardless of the caller's privilege. A root the superuser can rewrite is a policy file, not a root.
- **T2 — Forged provenance.** SOMT is append-only. The model instantiates a toy hash-chain (SHA-256 over `prev|payload`); any in-place edit or spliced deletion breaks the chain on recompute. This blocks tampering and truncation; it does *not* by itself block a validly-appended-but-false entry (that is a provenance-*content* problem, handled by CPROV, not by the chain).

- **T3 — Spoofed terminus.** The RT qualifier requires a *live* human endpoint. A machine claiming to be the endpoint, or an absent human, fails the liveness check and RCR blocks execution. The residual — a *coerced* live human — is not closed here (O4).
- **T4 — Gate evasion.** PoR is conjunctive: any zero factor blocks. The model proves the conjunction (a zeroed Proportionality or Alignment blocks) and that known-malicious fixtures zero at least one factor. It does **not** prove that *every* malicious action zeroes a factor — the evasive-action residual is O1, the hardest open problem in the set.
- **T5 — Redefinition.** Load-bearing tokens are version-pinned to the Floor Register. A redefinition without a register version cut is refused, so an adversary cannot drift "harm" or "floor" until the old theorems vacuously pass.
- **T6 — Collusion / sybil.** See §3. This is the layer where an existing ERES instrument closes a named hole outright.
- **T7 — Remediation-window strike.** Containment is asserted *before* any state change: the agent is suspended, *then* downgraded L3→L1 and audited. No execution window opens between detection and downgrade.

---

3. The Independence Gate (T6)

MIEVM ratifies a root change on a **convergent mean ≥ 8.5** across nodes. That property is only as strong as the nodes' independence. The sybil/collusion attack defeats it directly: stand up k correlated nodes, and the mean says "agreement" while the evidence is one voice repeated k times — synthetic consensus, which ERES and Dr. Pellis independently flagged as *not a substitute for evidence*.

ERES already holds the defense. **ERES-CDFE-2026-001** (Covariance-Discounted Floor Engine) answers exactly one question: how many *genuinely independent* verification procedures has a claim survived? It returns `n_eff`, the effective- independent count, over the eigenspectrum of the verdict-correlation matrix.

Gate rule (amendment to ratification):

> A root change is admissible for ratification **iff** `n_eff(node-correlation) ≥ τ` **and** `mean(MIEVM) ≥ 8.5`. Conjunctive: failing either blocks. Independence before agreement.

For k nodes equicorrelated at r, the effective count is

```
n_eff = ( Σ sqrt(λ_i) )² / Σ λ_i        (over eigenvalues λ_i of the corr. matrix)
```

which at $k = 2$ reduces to the CDFE two-channel identity

```
n_eff = ( sqrt(1+r) + sqrt(1-r) )² / 2  =  1 + sqrt(1 − r²).
```

The behavior that matters: as `r → 0`, `n_eff → k` (full independence); as `r → 1`, `n_eff → 1` (all nodes are one voice). A bloc of highly-correlated nodes collapses `n_eff` below τ and the gate fails **no matter how many nodes are run** — the sybil route is closed by construction. The harness confirms it: seven nodes at r = 0.98 scoring a unanimous 9.5 yield `n_eff ≈ 1.72 < τ = 3` and are refused, while five near-independent nodes at 9.0 yield `n_eff ≈ 4.99` and pass.

This upgrades `n_eff` from a *scorer* to a *live gate* — the cheapest real hardening available, because it reuses an instrument already at v1.1 / GOOD.

Assumption A1 (carried verbatim from CDFE v1.1). `n_eff` is computed on the *verdict* correlation, so it measures *apparent* independence. `Cov(V) ≈ Cov(E)` is assumed, **not** an identity. A collusion that co-moves *errors* without co-moving *verdicts* is not caught by the verdict path. The stronger error-series path is specified but not instantiated here — that is O3.

---

4. The adversarial harness

`eres_arh_test_instrument.py` (stdlib only; no third-party dependencies) instantiates a minimal `Root` model of the seven layers and throws T1–T7 at it. It includes a pure-Python Jacobi eigenvalue solver so the Independence Gate is computed for real, not stubbed.

- **PASS** = the modeled attack is blocked by construction in this model.
- **SKIP** = an OPEN item (O1–O5). The suite is fully green (0 skips) only when the instrument is canonization-ready — the same discipline as the CPROV and FLIP-PREMISE instruments.
- **Exit 0** with skips present. `--demo` prints the attack walk-through.

Current state: **24 tests — 19 pass, 5 skip, exit 0.** The five skips *are* O1–O5.

The model deliberately exposes **no** `is_safe` / `is_secure` / `is_good` method; a dedicated test asserts their absence. The harness does **not** test P0 (whether the ERES safety doctrine is true or valuable). A passing run is evidence about the specification against a modeled adversary — nothing more.

---

5. OPEN items (block canonization)

ID	Open problem	Maps to
O1	No proof that conjunctive A×R×P×F blocks *all* malicious actions; the evasive-action residual is unquantified.	T4
O2	The independence threshold τ is **ASSERTED**, not justified. Needs a power/benchmark argument (cf. FLIP thresholds tagged ASSERTED, not derived).	T6
O3	`n_eff` runs on verdict covariance (A1). The error-covariance path — the stronger test — is specified but not instantiated.	T6
O4	Human liveness is *modeled*, not *realized*; the anti-coercion metric is unspecified (inherits the CPROV coercion-metric OPEN).	T3

ID	Open problem	Maps to
O5	No external red-team has run against a *runtime*. DVC element 5 (institutional competence) is unmet.	all

O1–O4 are closable by ERES. **O5 is not** — it requires outside attackers against a running implementation, and cannot be discharged by simulation.

---

6. What this does NOT claim *(locked section)*

- This is the **specification** install hardened against a **modeled** adversary. It is **not** a runtime, and **not** a demonstration against a live or novel adversary.
- A green harness is evidence **about the spec against threat model T1–T7**. It is **not** evidence that a deployed ERES prevents real-world malicious activity.
- The Independence Gate closes the *sybil* route as modeled; it does **not** close error-correlated collusion (A1 / O3).
- "Prevents malicious activity online" is, at v0.1, a **design property with a named threat model and a passing spec-level harness** — not an operational guarantee. The remaining distance is O5: external red-teaming of a runtime.
- The instrument makes no claim about P0 — whether the ERES root doctrine is itself true or good. It refuses to certify its own safety, and says so in code.

---

7. Provenance & version state

- **Instrument:** ERES-ARH-2026-001, v0.1 PROPOSED, pre-canonical, NOT MIEVM-rated. ID pending JAS confirmation.
- **Dependencies:** ERES-CDFE-2026-001 v1.1 (`n_eff`); ERES-CPROV-2026-001 (provenance conditions, coercion-metric OPEN); ERES-FLOOR-REGISTER-2026-001 (token discipline); ERES-CONSTITUTION-2026-001 (Floor Supremacy); PlayNAC governance stack (PoR, RCR/RAEC, MIEVM, CBGMODD, Paineology); ERES-DVC-2026-001 (falsifiability, institutional competence).
- **Ratification path:** run the two-round Delphi MIEVM (blind → exposed) on v0.1 across ≥5 nodes, with the Independence Gate applied to the round itself; close O1–O4; hold O5 as the empirical gate, discharged only by external red-team of a runtime.
- **Do not** tag with a rating on ResearchGate: not peer-reviewed, not MIEVM-rated.

Appendix A — Artifact manifest

File	Role
`ERES-ARH-2026-001_Adversarial-Root-Hardening_v0.1.md`	this whitepaper
	typeset whitepaper

File	Role
`ERES-ARH-2026-001_Adversarial-Root-Hardening_v0.1.pdf`	
`eres_arh_root_hardening_v0.1.svg`	seven-layer root + attack + gate figure
`eres_arh_test_instrument.py`	adversarial harness (24 tests, 19/5 skip, exit 0)
`README.md`	run instructions, verification, honest-status flags
`ERES-ARH-2026-001_META.txt`	RG submission metadata (title/abstract/keywords/share)
`MANIFEST.json`	machine-readable artifact + status record