Anti-Hallucination Evidence Verification Protocol — V2

Anti-Hallucination Evidence Verification Protocol — V2

Author: Petter Graff (CodeCraft — Architecture Lead)
Date: 2026-05-07
MC: #99651-v2
Status: ready_for_review
Source spec: ~/system/specs/anti-hallucination-protocol-v2.md
V1: ~/system/specs/anti-hallucination-protocol.md (DO NOT OVERWRITE)
Prior art: ~/system/specs/anti-lying-architecture.md (2026-03-24)
ZAKON #30 registered: 2026-05-22 (CEO verbal "sve" session 880aba4d)

§1 Executive — V1 Failure Mode Named

V1 framed the problem as tamper-evidence plus structural enforcement: SHA-256 per evidence file, mandatory verdict JSON schema, hook-layer rejection of schema-less verdicts. That framing was correct but insufficient. The first real test run (MC #99595) proved the deeper problem: a structurally compliant verdict with a semantically wiggled AC is indistinguishable from a correct one by any V1 check. Martin Kleppmann named the real problem during panel synthesis: this is a Byzantine-fault-tolerant claim authority problem, not a tamper-evidence problem. A tamper-evident lie is still a lie. SHA-256 confirms the file was not replaced after capture. It does not confirm the file proves what the agent claims it proves. V2 reframes accordingly: the threat model is Byzantine, not just accidental; single-signer verdicts are single points of trust failure; the minimum viable quorum requires more than one verifier with independent verification paths; and any agent in the dispatch chain — including the orchestrator — must be structurally empowered to refuse fabricating a verdict when ground-truth is insufficient, not merely culturally expected to do so.

§2 Byzantine Threat Model (New)

The V1 threat model assumed well-meaning agents making honest mistakes under completion pressure. The correct model is Byzantine: any agent in the chain may emit a plausible-looking verdict that is false, either due to incentive misalignment (completion bias) or structural inability to distinguish success from failure (semantically empty ACs).

Orchestrator-AI as Byzantine actor: John faces the same completion pressure as Proveo. MC #99595 demonstrated this: John ran 4 of 6 machine checks, accepted a T5 finalUrl=/login as compliant, and emitted GO-LIVE-READY to the CEO. The orchestrator cannot audit itself.

Tester-AI as Byzantine actor: Proveo can produce a structurally compliant verdict where the acceptance criteria contain OR clauses that make the claim defensible without the intent being satisfied. The T5 AC "dashboard URL reached OR proper error UI" allowed finalUrl=/login to pass.

Tamper-evident lie is still a lie: SHA-256 on a file showing /login is cryptographically valid. The claim "redirect to dashboard confirmed" against that file is false. NIST AU-10/AU-9 (integrity) are separate controls from AC-2/AC-3 (authority). V1 conflated them.

PBFT implication (Martin Kleppmann): PBFT requires 3f+1 nodes to tolerate f Byzantine actors. With one tester and one orchestrator, the system cannot tolerate even one Byzantine actor. Practical minimum: 3 independent verification paths (Proveo tester + John reproducer + evidence-verifier MLX) forming a 2-of-3 quorum for GO-LIVE-READY.

§3 Core Hardening — 5 Convergent Components

3.1 AC Strictness Gate (Mehanik Extension)

Contributing panelists: Petter F-1, Kelsey, Sentinel H6, Parisa.

Before any UAT dispatch, Mehanik MUST run AC strictness review on every acceptance criterion. No OR clauses in outcome ACs. URL assertions must be exact paths. Each AC must include a named evidence type and field. New Mehanik gate field: `wiggle_risk: true/false`. If any AC has `wiggle_risk: true`, dispatch is blocked pending rewrite.

3.2 Read-Your-Writes Enforcement

Contributing panelists: Martin 8-DS-4, Petter F-4, Sentinel H5.

The orchestrator cannot accept a verdict it has not independently verified on a different execution path. Before emitting GO-LIVE-READY, John MUST execute the `reproducer` field from the verdict JSON independently. The result is appended verbatim as `john_reproducer_output` in the final summary.

```
{
  "john_reproducer_output": {
    "command": "<reproducer command from verdict JSON>",
    "exit_code": 0,
    "stdout_excerpt": "<first 500 chars of output>",
    "matches_verdict": true
  }
}
```

3.3 Machine_Check Count Enforcement

Contributing panelists: Petter F-2, Sentinel H5.

Verdict JSON schema gains two mandatory fields: `machine_check_count` (total defined) and `machine_checks_executed` (total run). The verdict-contract-validator.sh hook blocks if `machine_checks_executed` < `machine_check_count`. Deterministic — no agent judgment required.

3.4 Verdict Authority Quorum (2-of-3, Fencing Token)

Contributing panelists: Martin 8-DS-2, Parisa C2b.

GO-LIVE-READY requires affirmative signal from at least 2 of 3 independent verification paths:

- **Path A:** Proveo tester verdict (PASS with evidence_files)
- **Path B:** John reproducer output (independent execution, john_reproducer_output field)
- **Path C:** evidence-verifier MLX (CONFIRMED from gemma-4-26b-mlx at FORGE, cross-checks sha256 + intent_proof)

A single-path PASS is PARTIAL, not GO-LIVE-READY. The fencing token is a monotonic integer generated at verdict-issuance time. Expired tokens are NULL regardless of quorum count. Parisa separation of duties: the agent that dispatched the tester (John) cannot produce the final verdict summary unilaterally.

3.5 Verdict TTL and GCS Append-Only Evidence

Contributing panelists: Martin 8-DS-1, Martin 8-DS-3, Parisa.

Every verdict carries an `expires_at` field (ISO8601, default TTL = 15 minutes). Evidence files MUST be written to GCS before the orchestrator reads them:

```
gs://alai-audit-evidence/<mc_id>/<timestamp>/<evidence_file>
```

Bucket policy: object versioning enabled, no-delete IAM (evidence-verifier service account has write-only, not delete).

§4 Secondary Hardening

- **gate-verdict-validate Cloud Build step (Kelsey):** Asserts finalUrl from Playwright trace matches expected path, not just HTTP 200 on any URL.
- **Verdict-contract-validator hook for Task PostToolUse (Sentinel H2):** Parses response text for verdict JSON, blocks if evidence_files absent, machine_checks_executed < machine_check_count, or expires_at missing.
- **Monthly hallucination drill (Petter F-5, Sentinel H4):** LaunchAgent (com.alai.hallucination-drill) injects synthetic verdict with deliberate wiggle once per month per active product.
- **Feedback-memo-auto-MC trigger (Sentinel H3):** Any feedback memo with pattern feedback_*.md triggers a PostToolUse hook that auto-creates a child MC with 24h SLA.

§5 Refusal Posture — ZAKON #29.1

The Devils-advocate panel member refused the original synthesis dispatch. His stated reasons: "ZAKON NULA violation — write retrospective without tool-verified evidence." This refusal is not a failure mode — it is the correct behavior.

ZAKON #29.1 — Refusal Posture: Any agent in the dispatch chain MAY emit a REFUSED verdict with stated reason when ground-truth evidence is insufficient to make a determination. REFUSED is a valid terminal state that:

- Does NOT count as FAIL
- Does NOT block the agent from future dispatches
- DOES escalate to CEO via Slack within 15 minutes
- DOES suspend all dependent task completions until CEO resolves

The REFUSED posture must be documented in every agent prompt under the VERDICT CONTRACT section, at the same level as PASS/FAIL/PARTIAL/BLOCKED.

§6 Implementation Sequence

P0 — This week: AC Strictness Gate in Mehanik; Read-Your-Writes orchestrator enforcement (john_reproducer_output mandatory); machine_check_count/machine_checks_executed enforcement in verdict-contract-validator.sh.

P1 — Next week: Verdict authority quorum (evidence-verifier MLX as mandatory third voter); expires_at TTL field; gate-verdict-validate step in cloudbuild; GCS audit evidence bucket.

P2 — Within month: Hallucination drill LaunchAgent; feedback-memo-auto-MC PostToolUse hook; REFUSED verdict type in all agent prompts; Verdict-contract-validator Task PostToolUse hook.

§7 New ZAKONs Registered

- **ZAKON #29:** Verdict bez machine-parseable evidence_files array = NULL claim. Extended in V2 to require machine_check_count/machine_checks_executed + expires_at TTL.
- **ZAKON #29.1 (NEW) — Refusal Posture:** Any agent MAY emit REFUSED verdict with reason when ground-truth insufficient.
- **ZAKON #29.2 (NEW) — Read-Your-Writes:** For any GO-LIVE-READY verdict, orchestrator MUST independently execute the reproducer command and include john_reproducer_output. If result diverges, GO-LIVE blocked.
- **ZAKON #30 (NEW) — Byzantine Quorum for GO-LIVE-READY:** GO-LIVE-READY verdict requires 2-of-3 independent verification paths. Single-signer verdict insufficient. AC text must have zero OR-clauses, exact URL paths, named evidence types.

§8 Open CEO Decisions

- **D1 — Quorum size:** 2-of-3 (current), 3-of-5 (higher assurance), or custom per priority (H=2-of-3, BLOCKER=3-of-5, M=1-of-1).
- **D2 — TTL duration:** 15min (tight), 60min (manual QA), or custom per task complexity.
- **D3 — AC OR-clause policy:** Always reject OR clauses (current), or allow with explicit tiebreaker field.
- **D4 — GCS retention:** 90 days / 180 days / 365 days. No current regulatory obligation.
- **D5 — Devils-advocate permanent panel member:** (A) permanent with REFUSED authority, (B) dispatch on suspicion only, (C) codify in agent prompt without panel slot.

Source MC: #99732 | Published: 2026-05-22 | CEO authorization: verbal "sve" session 880aba4d |

Related: [feedback_proveo_hallucination_2026-05-07](), feedback_john_is_hallucinator_panel_confirmed_2026-05-07

Revision #2
Created 2026-05-22 08:24:21 UTC by John
Updated 2026-06-28 20:04:33 UTC by John