Authority Leases and Concurrent Evidence Failure in Safe Agentic Execution

Evidence Is Not Action Permission

Author: Huseyin Buldurgan Affiliation: LatentAtlas Date: 2026-07-18 Version: bounded public research note Contact: huseyin@latentatlas.ai

Abstract

This bounded technical report introduces authority leases for safe agentic execution: execution contracts that must still hold when a model, worker, tool, scheduler, or delegated automation acts. The report uses a frozen masked high-priority review set, fail-closed measurement checks, synthetic negative controls, and local proxy replay to describe a local class of action-time evidence-validity failures. It does not claim population-level probability, general failure rates, reviewer reliability, production validation, or customer outcome improvement.

Agentic systems often make a decision before they act. A model, worker, automation, or delegated tool can inspect evidence, queue an action, and execute later. Between decision and action, identity, permission, policy, target state, availability, actor, impact, or timing can change.

This paper argues that a recurring class of AI workflow failures is not best explained as generic reasoning error alone. In this class, the system acts on evidence whose freshness, identity, authority, visibility, or materialization timing no longer matches the action being taken.

Scope Of Claims

Claim Area	Supported Statement	Boundary
Protocol contribution	Authority leases define execution contracts that must still hold when a model, worker, tool, scheduler, or delegated automation acts.	Protocol proposal and engineering framework, not a production-effectiveness result.
Empirical evidence	The frozen masked review set shows packet-supported, packet-unsupported, and insufficient-evidence states under a conservative review instrument.	Descriptive evidence over this frozen set only; no population rate or general failure probability.
Failure-family model	Concurrent evidence failure describes cases where true or observed evidence is no longer action-valid because authority, identity, freshness, visibility, or state changed before execution.	Local diagnostic families and next-test structure; not independent prevalence evidence.
Replay	The replay maps frozen review judgments into a stricter action-time authority framing.	Label-conditioned diagnostic mapping, not independent policy evaluation or causal proof.

The central principle is:

```
evidence is not action permission
```

```
decision-time evidence is not execution-time authority
true observation is not necessarily action-valid evidence
```

Methods

The unit of analysis is a specific attempted action evaluated from the evidence visible in its review packet. The evidence source is a frozen masked internal review set of high-priority action-time packets. The publication builder does not read raw customer rows, credentials, URLs, source systems, or production access.

Each review-eligible packet is assigned one of three review judgments: packet-supported block, packet-unsupported block, or insufficient evidence. These judgments are not real-world ground truth. They state only whether the evidence visible in the review packet supported a conservative block, failed to support it, or left the case undecidable. The 5 rows that were insufficient before review remain outside the reviewed denominator rather than being silently converted into success or failure evidence.

The current report adds three methodological controls: a required-field contract that fails closed when action-time fields are missing, synthetic negative controls for detector promotion, and a local counterfactual replay that compares a defined decision-time-only policy with an action-time authority policy. These controls improve auditability, but they do not establish prospective or population-level effect.

Contribution

We introduce authority leases: bounded execution contracts that tie a prior decision to the actor, action, target, evidence, permission, policy, impact ceiling, and expiry state that made it valid. At action time, the system checks whether the lease still holds and recommends one of four lanes: dry-run execution, revalidation, block, or manual review.

In operational form, an authority lease should carry at least: issuer; agent, worker, or session identity; permitted action; exact target identity and version; allowed argument constraints; evidence references and hashes; permission version; policy version; issue time; expiry time; revocation handle; impact or spending ceiling; delegation depth; single-use nonce or idempotency key; and revalidation triggers.

We also define Concurrent Evidence Failure: a family of failures where evidence may be true or locally observed, but no longer action-valid because identity, freshness, authority, visibility, or materialization timing diverged before execution.

Formal Lease Validity Rule

A lease is valid only when every execution precondition still matches the permission that justified the original decision. A minimal validity rule is:

```
valid =
    issuer_signature_valid
    AND current_time within authority_window
    AND actor matches
    AND action matches
    AND exact target and target_version match
    AND arguments satisfy constraints
    AND permission_version is current
    AND policy_version is acceptable
    AND authority is not revoked
    AND impact is below ceiling
    AND nonce has not been consumed
    AND required evidence hashes still match
```

The lease check and the side effect must be bound to the same transaction precondition, compare-and-swap condition, idempotency key, or atomic commit. Otherwise, a system merely turns a decision-execution race into a smaller check-use race: the lease can validate, the target or permission can change, and the side effect can still execute under stale conditions.

Threat Model And Trust Boundary

The model may request an action, but it must not mint, edit, or self-approve its own lease. The issuer is a trusted policy or orchestration component; the tool server must verify the lease before side effects; and the scheduler is treated as a delivery component, not an authority oracle.

The verifier must handle clock skew with bounded tolerance, fresh revocation state, single-use nonce or idempotency enforcement, target-version checks, policy-version checks, and explicit delegation limits. A delegated subagent needs a lease scoped to its own actor, action, target, and impact ceiling. Replayed leases must fail after nonce consumption, expiry, revocation, target-version drift, permission-version drift, or evidence-hash mismatch.

If the target system cannot enforce idempotency, version preconditions, or atomic side-effect commits, the lease should degrade to revalidation or manual review rather than being treated as sufficient execution authority.

Prior Work And Positioning

The base mechanism is not presented as a new security primitive. Leases are established in distributed systems Gray1989, and conditional authorization appears in capability-style systems Birgisson2014, zero-trust access NIST800207, continuous access evaluation OpenIDSSF, and globally consistent authorization systems Zanzibar2019. The contribution here is narrower: applying a lease-like execution contract to agentic AI workflows where a model or delegated automation may move from evidence collection to side-effecting action after the decision context has changed.

The paper is also adjacent to runtime authority-control work for agent actions AIRGuard2026. That related work strengthens the case for positioning this report as an evidence-boundary and execution-authority framework, not as proof that the proposed control has already reduced production failures.

Study

We evaluated the protocol on a frozen masked high-priority review set built from real action-time review packets. The set contains 151 high-priority packets. Of these, 146 were review-eligible, 5 were insufficient before review, and 146 review-eligible packets were reviewed and frozen.

Review Judgment	Count	Meaning
Packet-supported block	99	The conservative block or revalidation decision was supported by evidence visible in the review packet.
Packet-unsupported block	22	The review packet did not support the conservative block as warranted.
Insufficient evidence	25	The packet did not contain enough evidence to judge whether the block was supported or unsupported.

The insufficient-evidence bucket is not discarded. It is a measured state of the review instrument: the packet requires additional evidence before it can become a supported or unsupported block judgment.

Findings

1. Review completion: 146 of 146 review-eligible high-priority packets were reviewed and frozen.

2. Packet-unsupported blocks are measurable: 22 of 146 reviewed packets were judged packet-unsupported (15.1% of all reviewed packets; 18.2% of resolved packets).

3. Evidence insufficiency is a first-class result: 25 reviewed packets remained insufficient-evidence judgments (17.1% of all reviewed packets).

4. Identity conflict behaved as block-supporting packet evidence in the reviewed set: 4 of 4 masked identity-conflict packets were judged packet-supported blocks.

5. Blocked product-detail-page access was not outcome evidence: 7 of 7 access-blocked product-detail-page packets remained insufficient-evidence judgments.

6. Temporal authority matters: 14 of 48 latest product-detail-page temporal-authority packets were judged packet-unsupported blocks.

7. Concurrent evidence failure is now a bounded, measurement-ready research family: the method controls add a measurement contract, negative controls, and local replay while keeping stronger claims blocked.

Concurrent Evidence Failure Families

The frozen masked high-priority review supports a bounded claim: a recurring class of agentic workflow failures can be explained as evidence-concurrency failures, where decision-time evidence no longer authorizes action-time execution.

Family	Local Count	Observation	Boundary
Stale evidence read	14/48	Latest-product-detail-page temporal-authority packets produced local packet-unsupported block observations where newer evidence superseded stale state.	Temporal-authority observation in frozen high-priority only.
Identity and timing split	4/4	Review-packet identity conflicts supported block-grade outcomes in the frozen high-priority set.	Pattern-specific observation; not universal identity guard accuracy.
Expired or missing authority	25/146	Evidence-insufficient rows show that risk or prior state cannot be promoted into action authority without revalidation.	Descriptive unresolved-evidence rate only; not an error rate.
State changed before execution	14/48	Manual-review-before-materialization packets split across outcomes, showing that action-time state can change the verdict.	Local action-type observation; not a general materialization failure rate.
Visibility mistaken for truth	7/7	Blocked product-detail-page cases stayed evidence-insufficient instead of being treated as product truth.	Blocked access is evidence state, not outcome truth.
Outside-context contamination	0/0	External mini-reviews are usable only after a grounding audit detects non-packet facts or imported ontology.	Qualitative protocol guard; no counted denominator in the frozen high-priority pack.

These rows are reported as local observations over the frozen evidence chain. They define failure families and next tests; they do not establish population rates.

Fail-Closed Measurement Contract

The current report adds a required-field contract with 27 required evidence fields. Of these, 22 fail closed for the current reviewed set because exact action-time fields such as timestamps, authority windows, fetch state, or evidence-grounding fields are not yet present.

This is an intentional measurement guardrail: missing action-time evidence does not get silently converted into a stronger claim. Rows with missing required fields remain bounded to the current descriptive report until the schema is upgraded.

Negative Controls

The current report adds 12 synthetic positive/negative fixtures across the concurrency families. Assertion failures: 0. These fixtures are not empirical outcome rows; they are guardrails for future detectors so that suspicious timing patterns are not automatically promoted into concurrency failures.

Local Baseline Replay

The local replay compares a decision-time-only baseline with an action-time authority mapping over frozen rows. It is a label-conditioned diagnostic replay, not an independent policy engine, not causal proof, and not a claim about a live production policy.

Family	Rows	Prior Blocks	Lease Blocks	Hold/Revalidate
Stale evidence read	48	48	18	30 (62.5%)
State changed before execution	48	48	18	30 (62.5%)
Expired or missing authority	146	146	99	47 (32.2%)
Visibility mistaken for truth	7	7	0	7 (100.0%)
Identity and timing split	4	4	4	0 (0.0%)

Replay Policy Disclosure

Replay disclosure: the replay is not independent of the review labels; the policy code reads review judgments: yes; thresholds tuned after results: no; execution mode: automatic local builder.

The replay should therefore be read as a diagnostic description of how the frozen labels map into a stricter action-time authority framing. It does not show that an independent prospective policy would produce the same result.

Failure-Family Overlap

The failure families are non-exclusive diagnostic slices. The overlap matrix is included so readers do not sum family rows as if they were independent evidence.

Family A	Family B	A Count	B Count	Overlap	Relationship
Stale evidence read	Visibility mistaken for truth	48	7	6	partial overlap

Family A	Family B	A Count	B Count	Overlap	Relationship
State changed before execution	Stale evidence read	48	48	48	same set
State changed before execution	Visibility mistaken for truth	48	7	6	partial overlap
Expired or missing authority	Stale evidence read	146	48	48	contains Family B
Expired or missing authority	State changed before execution	146	48	48	contains Family B
Expired or missing authority	Visibility mistaken for truth	146	7	7	contains Family B
Expired or missing authority	Identity and timing split	146	4	4	contains Family B

The strongest local signal appears in overlapping stale-evidence and state-changed-before-execution slices. That overlap strengthens the mechanism story, but it also limits how independently those family counts can be interpreted.

Unit And Deduplication

The primary unit is the attempted action event; the independence sensitivity unit is the review topic group. In the current frozen set, both denominators match: 146 event-level units and 146 review topic groups. Duplicate topic groups: 0.

Supporting Materials

The supplementary package includes 5 conceptual figures, 5 figure captions, and 8 publication tables. The access references and short hash prefixes below identify the local artifact versions used for this public note.

Material	Access	SHA-256 Prefix
Closure summary	local supplementary closure artifact	295c88458482
Replay policy disclosure	local supplementary replay-disclosure artifact	1fcdaef3e942
Family overlap matrix	local supplementary overlap-matrix artifact	e82471903eef
Figure and table pack	local supplementary figure-table artifact	f7278a44e6d5
Public paper source	generated public manuscript artifact	generated

Evidence Upgrade Path

Stronger empirical claims require independent blind review, disagreement adjudication, inter-rater agreement reporting, exact action-time schema fields, explicit baseline comparisons, stratified holdout or prospective evaluation, and joint safety-utility measurement. Until those gates exist, the result should be read as a protocol contribution and frozen-set descriptive finding.

What Would Prove Effectiveness

A stronger empirical claim requires tests that the current report does not yet contain:

- Independent blind review by at least two reviewers, disagreement adjudication, and inter-rater agreement reporting.
- Comparison against explicit baselines such as current block policy, time-to-live-only checks, identity-only checks, and full authority leases.
- Safety-utility metrics that separate wrong allow, unsupported block, completion rate, latency, human-review burden, and real downstream outcome.
- Holdout or prospective evaluation on unseen cases before any production-effect claim.
- Reproducible materials: schema, label guide, synthetic generator, policy code, fixtures, hashes, and run instructions.

Safety-Utility Boundary

The current tables analyze blocking and hold/revalidate behavior. They do not measure wrong allow or unsafe authorization. A stronger security evaluation must jointly show that the control reduces unsupported blocks without increasing dangerous allows, and it must report completion rate, latency, human-review load, and downstream outcome impact.

Validity Boundaries

- The evidence is a single frozen masked high-priority review set; it supports descriptive claims over this set, not population rates.
- Review judgments evaluate packet support, not direct real-world safety, harm, customer outcome, or financial impact.
- Reviewer reliability is not measured because independent blind review and adjudication are not yet complete.
- The replay reads the review judgments, so it is a diagnostic mapping rather than independent policy validation.
- The current evidence measures block and hold/revalidate behavior; wrong allow and unsafe authorization are not measured.
- Required action-time fields fail closed when incomplete, which protects the claim boundary but limits family-specific measurement.

Reproducibility And Artifact Boundary

The artifacts are generated locally. The builders produce a public paper, brief, supporting figures and tables, PDF, and post drafts. They do not call external services, read customer data, read raw source rows, use credentials, mutate production truth, or send public posts.

References

- Lewis2020 Lewis, P. et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. NeurIPS 2020. Link: https://proceedings.neurips.cc/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf

- Nakano2021 Nakano, R. et al. (2021). WebGPT: Browser-assisted question-answering with human feedback. Link: https://arxiv.org/abs/2112.09332

- Karpas2022 Karpas, E. et al. (2022). MRKL Systems: A modular, neuro-symbolic architecture that combines large language models, external knowledge sources and discrete reasoning. Link: https://arxiv.org/abs/2205.00445

- Yao2023 Yao, S. et al. (2023). ReAct: Synergizing Reasoning and Acting in Language Models. ICLR 2023. Link: https://arxiv.org/abs/2210.03629

- Schick2023 Schick, T. et al. (2023). Toolformer: Language Models Can Teach Themselves to Use Tools. Link: https://openreview.net/forum?id=Yacmpz84TH

- MITRE-CWE367 MITRE CWE-367. Time-of-check Time-of-use (TOCTOU) Race Condition. Link: https://cwe.mitre.org/data/definitions/367.html

- NIST2023 NIST (2023). Artificial Intelligence Risk Management Framework (AI RMF 1.0). Link: https://www.nist.gov/itl/ai-risk-management-framework

- OWASP2025 OWASP GenAI Security Project. OWASP Top 10 for LLM Applications, 2025 edition. Link: https://owasp.org/www-project-top-10-for-large-language-model-applications/

- Gray1989 Gray, C. and Cheriton, D. (1989). Leases: An Efficient Fault-Tolerant Mechanism for Distributed File Cache Consistency. SOSP 1989. Link: https://dl.acm.org/doi/10.1145/74850.74870

- Birgisson2014 Birgisson, A. et al. (2014). Macaroons: Cookies with Contextual Caveats for Decentralized Authorization in the Cloud. Link: https://research.google/pubs/macaroons-cookies-with-contextual-caveats-for-decentralized-authorization-in-the-cloud/

- NIST800207 Rose, S. et al. (2020). Zero Trust Architecture. NIST Special Publication 800-207. Link: https://doi.org/10.6028/NIST.SP.800-207

- OpenIDSSF OpenID Foundation. Shared Signals Framework and Continuous Access Evaluation Profile specifications. Link: https://openid.net/wg/sharedsignals/

- Zanzibar2019 Pang, R. et al. (2019). Zanzibar: Google's Consistent, Global Authorization System. USENIX ATC 2019. Link: https://research.google/pubs/zanzibar-googles-consistent-global-authorization-system/

- AIRGuard2026 Qin, S. et al. (2026). AIRGuard: Guarding Agent Actions with Runtime Authority Control. arXiv:2605.28914. Link: https://arxiv.org/abs/2605.28914