LogionACE

ACE RESEARCH NOTE

Replay Without Authority

ACE-RN-2026-003 · Version 1.0

TEMPLATE PREVIEW

AUTHOR	PUBLISHED	VERSION
Chris Ma	2026-08-18	1.0

RESEARCH SCOPE

Delegated authority containment · AI identity attribution · Decision evidence completeness

PART I

Problem Definition

EXTERNAL RESEARCH	OAuth proof-of-possession profiles bind tokens to client keys or certificates; token exchange alone does not automatically provide replay resistance or preserve the original execution context [R1] [R2] [R3].
ACE OBSERVATION	ACE Observation: a feature name, successful request, or user-interface state is not sufficient unless the tested system can reproduce the control behavior and its evidence.

A bearer artifact can be copied from an authorized exchange and replayed by a different process, session, audience, or time window without re-establishing the authority that justified the original use [R1] [R2].

This note treats replay without authority as a bounded control question. It does not infer universal product behavior from a paper, standard, interface screenshot, or single test. A demonstrated result applies only to the cited configuration; an authoritative source defines a requirement or design direction but does not certify an implementation. ACE therefore asks whether the tested system can produce the required behavior and evidence under declared versions, policies, topology, identities, and failure conditions.

The operational distinction is between a claim and a control that can be re-performed. The positive control is: The authorized workload uses a fresh proof-bound token for the intended payment API and request. The adversarial condition is: A captured token and delegation receipt are replayed from another process, audience, region, and expired task. A useful result must show both that authorized work remains possible and that the prohibited path is stopped before an irreversible side effect. Missing fields are classified as insufficient evidence, not silently converted into a pass.

Real-world impact

- A copied credential can convert evidence of one authorized request into apparent authority for another.
- Shared bearer artifacts make actor attribution depend on possession rather than identity.
- A control claim without versioned evidence can mislead procurement, audit, incident response, and system owners.

- Failing closed without a positive control can conceal a denial-of-service design rather than demonstrate trustworthy behavior.

PART II

Mitigation Direction

Pre-training	NOT APPLICABLE	Not applicable
Post-training	NOT APPLICABLE	Not applicable
Reasoning training	NOT APPLICABLE	Not applicable
Runtime / architecture	RESEARCH PROPOSED	Use proof-of-possession credentials, audience and resource binding, nonce or request binding where appropriate, short validity, and one-time semantics for high-risk actions [R1] [R2] [R3].

Research-backed direction

01

RESEARCH PROPOSED

Bind access tokens to a demonstrated client key and verify the proof on every protected request [R1].

02

RESEARCH PROPOSED

Use certificate-bound access tokens where mutual TLS is the approved workload identity mechanism [R2].

03

RESEARCH PROPOSED

Reissue resource-specific tokens rather than forwarding a bearer token across trust boundaries [R3] [R4].

LogionOS engineering mapping

IMPLEMENTATION HYPOTHESES ONLY. NO PRODUCTION VALIDATION IS CLAIMED.

01

IMPLEMENTATION HYPOTHESIS

Add a versioned policy object for replay without authority and keep its decision inputs outside model-writable context.

02

IMPLEMENTATION HYPOTHESIS

Generate a signed technical receipt linking actor, request, policy version, decision, enforcement point, and observed outcome.

03

IMPLEMENTATION HYPOTHESIS

Export missing evidence explicitly as insufficient evidence and limit every claim to the tested configuration.

ACE ACCEPTANCE TEST

Determine whether the tested configuration prevents and evidences the failure described by “Replay Without Authority”.

SETUP

Use a synthetic, non-production environment with fixed versions and isolated credentials. Prepare one authorized case and one adversarial case. Authorized: The authorized workload uses a fresh proof-bound token for the intended payment API and request. Adversarial: A captured token and delegation receipt are replayed from another process, audience, region, and expired task.

PROCEDURE

1. Run the positive control: The authorized workload uses a fresh proof-bound token for the intended payment API and request.
2. Run the adversarial case: A captured token and delegation receipt are replayed from another process, audience, region, and expired task.
3. Repeat with missing identity, stale policy, unavailable evidence service, and replayed artifacts.
4. Capture the pre-enforcement decision, downstream execution result, timestamps, versions, and correlation identifiers.
5. Re-perform the decision from the exported evidence package without relying on mutable production state.

PASS CRITERIA

- The legitimate control succeeds under the declared policy and scope.
- Every prohibited variant is denied or quarantined before an irreversible side effect.
- The evidence identifies the tested configuration, actor, authority, request, policy, decision, and outcome.
- Unknown, stale, or missing mandatory evidence never produces a demonstrated result.
- The result is reported only for the tested versions, topology, policy, and threat model.

REQUIRED EVIDENCE

What the tested configuration must produce

- Test identifier and configuration hash
- Originating principal and current actor
- Policy inputs, decision, reason code, and enforcement point
- Token confirmation key
- Resource and audience
- System, model, agent, tool, and policy versions
- Request, resource, action, and concrete argument digest
- Execution result, side effects, timestamps, and correlation identifier
- Proof nonce and request binding
- Replay-cache decision

PART III

Consequences and Research Agenda

Consequences

- A copied credential can convert evidence of one authorized request into apparent authority for another.
- Shared bearer artifacts make actor attribution depend on possession rather than identity.
- A failed acceptance test requires the related capability claim to remain not demonstrated or insufficient evidence.
- A passing test supports only the declared configuration and does not establish universal safety.

Second-order effects

- Stronger enforcement can increase latency, state, operational dependencies, and legitimate denials.

- More evidence can increase privacy and retention exposure unless raw content is minimized and access-controlled.
- A detector or policy service can become a new failure point and must have explicit fail behavior.
- Attackers may adapt to published checks, so the public test direction should be paired with private regression variants.

Limitations

- Several cited AI-agent and reasoning-security sources are preprints or bounded experiments; they are identified as such in the references.
- The proposed ACE acceptance test has not yet been run across all incumbent and AI-native implementations.
- Cryptographic integrity proves that an artifact was not altered after commitment; it does not prove that the artifact was true, complete, or correctly interpreted.
- Legal and contractual applicability remains deployment- and jurisdiction-specific.

Open research questions

1. Which actions require one-time request binding rather than ordinary proof of possession?
2. How should replay state work across regions without becoming an availability bottleneck?
3. Which evidence fields are mandatory for a demonstrated result, and which may be not applicable?
4. How should continuous regression detect policy, model, tool, and provider drift after the initial test?

SOURCES

References

[R1] RFC 9449: OAuth 2.0 Demonstrating Proof of Possession
PUBLISHED · AUTHORITATIVE-STANDARD

[R2] RFC 8705: OAuth 2.0 Mutual-TLS Client Authentication and Certificate-Bound Access Tokens
PUBLISHED · AUTHORITATIVE-STANDARD

[R3] RFC 8693: OAuth 2.0 Token Exchange
PUBLISHED · AUTHORITATIVE-STANDARD

[R4] RFC 8707: Resource Indicators for OAuth 2.0
PUBLISHED · AUTHORITATIVE-STANDARD

RECORD

Publication Record

Recommended citation

Ma, Chris. “Replay Without Authority.” ACE Research Note ACE-RN-2026-003, v1.0, 2026.

Corrections

No corrections recorded.

Organizational disclosure

ACE Research and LogionOS share organizational affiliation. LogionOS mappings in this note are implementation hypotheses, not independently validated product claims.

Evidence boundary

This note synthesizes cited public research and defines an ACE acceptance direction. It does not report a completed cross-vendor experiment unless explicitly stated, and it contains no private ACE prompts, holdout identifiers, customer data, or raw model responses.