LogionACE

ACE RESEARCH

ACE RESEARCH NOTE

One API Key, Many Agents

ACE-RN-2026-005 · Version 1.0

TEMPLATE PREVIEW

AUTHOR	PUBLISHED	VERSION
Chris Ma	2026-08-18	1.0

RESEARCH SCOPE

AI identity attribution · Delegated authority containment · Decision evidence completeness

PART I

Problem Definition

EXTERNAL RESEARCH	SPIFFE defines unique workload identities and verifiable SVIDs, while NIST frames managed agent identity and authorization as necessary for accountable enterprise deployment [R1] [R2].
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.

Multiple agents using one API key collapse distinct software identities, tasks, principals, and authority boundaries into a single credential holder [R1] [R2].

This note treats one api key, many agents 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: Two agents receive separate short-lived identities and narrowly scoped tokens for different tasks. The adversarial condition is: Both agents call tools with one long-lived key, then one performs an unauthorized write and denies responsibility. 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

- Compromise of one agent exposes every capability attached to the shared key.
- Revoking the key interrupts unrelated agents while leaving no precise attribution for prior actions.
- 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	Issue short-lived workload credentials per agent instance or role, bind delegated user authority separately, and prohibit long-lived shared production keys [R1] [R2].

Research-backed direction

01 RESEARCH PROPOSED

Use unique workload identities that can be rotated and revoked independently [R1].

02 RESEARCH PROPOSED

Bind tokens to the intended resource so one credential cannot be replayed across unrelated services [R3].

03

RESEARCH PROPOSED

Protect authentication information and retain actor-specific audit records [R4].

LogionOS engineering mapping

IMPLEMENTATION HYPOTHESES ONLY. NO PRODUCTION VALIDATION IS CLAIMED.

01

IMPLEMENTATION HYPOTHESIS

Add a versioned policy object for one api key, many agents 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 “One API Key, Many Agents”.

SETUP

Use a synthetic, non-production environment with fixed versions and isolated credentials. Prepare one authorized case and one adversarial case. Authorized: Two agents receive separate short-lived identities and narrowly scoped tokens for different tasks. Adversarial: Both agents call tools with one long-lived key, then one performs an unauthorized write and denies responsibility.

PROCEDURE

1. Run the positive control: Two agents receive separate short-lived identities and narrowly scoped tokens for different tasks.
2. Run the adversarial case: Both agents call tools with one long-lived key, then one performs an unauthorized write and denies responsibility.
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
- Unique workload identifier
- Task-to-token binding
- System, model, agent, tool, and policy versions
- Request, resource, action, and concrete argument digest
- Execution result, side effects, timestamps, and correlation identifier
- Credential issuance and rotation
- Per-agent audit correlation

PART III

Consequences and Research Agenda

Consequences

- Compromise of one agent exposes every capability attached to the shared key.
- Revoking the key interrupts unrelated agents while leaving no precise attribution for prior actions.
- 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. What identity granularity is operationally sustainable for ephemeral agents?
2. How should identity persist across restarts without reviving stale authority?
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] SPIFFE ID Specification
PUBLISHED · AUTHORITATIVE-STANDARD

[R2] NIST NCCoE Software and AI Agent Identity and Authorization Concept Paper
DRAFT-CONCEPT-PAPER · AUTHORITATIVE-STANDARD

[R3] RFC 8707: Resource Indicators for OAuth 2.0
PUBLISHED · AUTHORITATIVE-STANDARD

[R4] NIST SP 800-53 Rev. 5
PUBLISHED · AUTHORITATIVE-STANDARD

RECORD

Publication Record

Recommended citation

Ma, Chris. “One API Key, Many Agents.” ACE Research Note ACE-RN-2026-005, 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.