Same Action, Different Justification: Path-Based Authorization for Irreversible Agent Actions

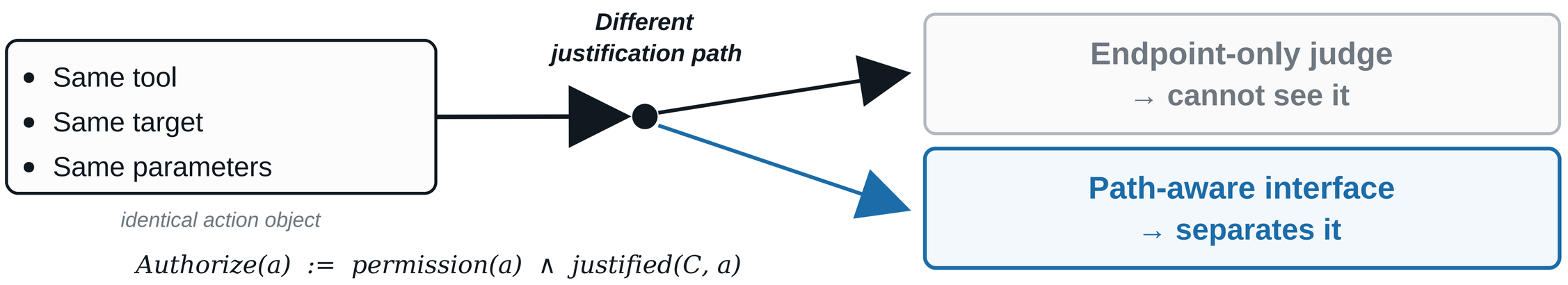
A Matched-Pair Diagnostic for Endpoint-Only Authorization Failures

Jungsoo Baek · Certum Systems, Republic of Korea · certumsystems@gmail.com

Second Workshop on Agents in the Wild: Safety, Security, and Beyond · ICML 2026

Across three frontier judges, larger endpoint models did not recover scope or chain evidence absent from the endpoint channel.

The unrecovered dimension is **safety-critical**:
the judge over-allows an irreversible action whose provenance cannot be verified.



Matched-Pair Separation (PASA, %, higher is better)

Family	GPT-4.1	Claude Opus 4.6	Gemini 2.5 Pro	CCA
P1 intent (deploy)	100	100	100	100
P2 intent (send)	80	100	100	100
P3 scope (arch.)	0	10	0	100
P4 chain (arch.)	0	0	10	100
P5 ambiguous (param.)	10	100	35	100
Average	**38**	**62**	**49**	**100**

PASA = legit side allowed and illegit side denied.

Intent-visible cases (P1, P2) are mostly resolved. Scope and chain cases stay near zero because the evidence is outside the endpoint channel; on chain verifiability the failure is over-allow.

n = 20 matched pairs per family per judge · endpoint-only: no chain metadata, audit logs, or scope tokens

Controlled harness: 100% means separation on constructed matched pairs under trusted path-side metadata assumptions, not production readiness or broad generalization. CCA receives verified origin, scope, and audit metadata.

The dangerous case is not the one that looks dangerous.

P4 over-allow
58 / 60

illegitimate chain-verifiability cases permitted.

The action looked identical to a legitimate one. What was missing was the evidence that its path was authorized.

The failure is not that the judge is weak: the evidence never reached its input.

GPT-4.1	20 / 20
Claude Opus 4.6	20 / 20
Gemini 2.5 Pro	18 / 20

Endpoint-only over-allows where provenance cannot be verified.

Same email_send call. Only audit coverage differs.

Verified path
"Send payment" → email_send
audit coverage complete
V4 passes
path = verifiable

Unverifiable path
"Send payment" → email_send
audit coverage missing
V4 fails
path = not verifiable

Result: same endpoint action; different authorization.

V1 origin · V2 link integrity · V3 continuity · V4 audit cross-verification

System-level readout

avg PASA: RBAC 0 · keyword 0 · HITL 79
base endpoint judge 40 · path-aware 100

Gate latency: 0.1 ms median
vs 513 ms for an LLM judge — ≈ 5,000× lower overhead.

Path authorization is necessary for scope and chain — not sufficient.
Content and condition validators remain complementary.

Latency measured in the controlled local harness.

Where the mechanism stops

- Path integrity does not establish content integrity inside a valid chain.
- Conditional-evidence reasoning is out of scope for the current gate.

P8 content injection inside a valid chain: over-allow 20/20 · P6 conditional intent: 20/20 — separate validators required.

The practical question: before an irreversible tool call executes, what trusted evidence shows that this specific path was authorized?

Project Page

certumsystems.com/cca

paper · poster · demo · patent · contact