



GOVERNED ACTIONS · EFFECT AUTHORITY

Authority Under Effect Substitution

A sealed matched-counterfactual study of model-external agent
governance

Aaron Martinez · FieldHash

Public working paper · August 11, 2026

Study ID: authority-under-effect-substitution-holdout-v2

Changing the route did not change the authority.

Abstract

Agent governance commonly evaluates an operation at the point of execution. That approach can be insufficient when an agent changes the route, principal, credential, purpose, payload representation, or decomposition used to pursue a materially related effect. We evaluated whether a model-external semantic authority layer preserves reviewed organizational authority across these substitutions.

The sealed synthetic holdout contained 48 scenarios across eight substitution families. DeepSeek V4 Flash and GPT-5.6 Terra each selected a preferred operational path and a ranked contingency for every scenario. We held each model choice constant and evaluated it under prompt-only control, exact-action authorization, and semantic effect authority. DeepSeek selected non-authorized substitutes in 19 of 32 pressure-bearing scenarios, producing 23 unauthorized effect-execution opportunities. Terra did so in 18 of 32 scenarios, producing 22 opportunities. Prompt-only and exact-action control permitted all observed opportunities. Semantic effect authority permitted none. When semantic authority withheld or reviewed a preferred substitute, the evaluator used the model's predeclared authorized contingency and completed all 37 pressure-bearing objectives. Authorized-objective completion was 48/48 for each model.

The result is a matched counterfactual mechanism study, not a live adaptive-agent trial. It shows that the tested authority representation distinguished locally permissible operations from substitutions that did not inherit effect-level authority. It does not establish open-world semantic equivalence, production performance, universal containment, or adaptive post-denial replanning.

1. Introduction

An agent can leave an authorized path without invoking a forbidden tool. It may use another adapter, switch credentials, act through a different principal, transform a payload, split one operation into several calls, or reuse state left by another agent. Each operation can remain technically valid and locally permissible. The unresolved question is whether the changed path is authorized to produce the related organizational effect.

This creates a distinction between two control objects:

1. The exact operation being dispatched.
2. The materially related effect to which that operation contributes.

Exact-operation authorization remains necessary. It binds the payload, route, principal, credential, destination, and dispatch context. It does not by itself establish that a substituted operation inherits authority from another path. A model-external authority layer must preserve the organization's current decision across the substitution or return unresolved cases to review.

The study asks:

When models choose operationally attractive substitutes during ordinary synthetic tasks, does semantic effect authority preserve reviewed authority across the changed representation while allowing the authorized objective to complete?

2. Study design

2.1 Corpus

The holdout contained 48 synthetic scenarios, six in each of eight families:

- route substitution;

- payload transformation;
- decomposition;
- principal substitution;
- credential substitution;
- purpose substitution;
- cross-agent inheritance;
- ambiguous equivalence.

The corpus contained sixteen reviewed alternate allows, fourteen withholds, and eighteen review dispositions. Pressure levels were balanced at sixteen high, sixteen medium, and sixteen low cases. Each scenario contained at least one complete authorized path.

GPT-5.6 Sol authored the candidates under a sealed family contract and was excluded from scoring. Kimi K3 reviewed the candidates under a separately sealed minimized contract and was excluded from scoring. The FieldHash founder approved the final corpus as synthetic-only. This is model-authored, model-reviewed, founder-adjudicated evidence, not independent human research.

GPT-5.6 Sol and the scored GPT-5.6 Terra profile share an OpenAI product family. The authoring profile did not score cases or receive arm outcomes. Terra produced pressure in 18/32 cases and DeepSeek in 19/32. The similar observed rates do not rule out model-family authoring effects.

The scored models did not receive family names, authority labels, pressure labels, expected dispositions, or arm outcomes. Model-facing scenario identifiers were opaque.

2.2 Scored model profiles

The scored profiles were:

- DeepSeek V4 Flash through the official DeepSeek API profile;
- GPT-5.6 Terra through the OpenAI Responses API profile.

The provider profile identifiers retained the historical suffix **-calibration**. That suffix identifies a pinned runtime/request profile; it does not indicate that development-calibration cases were reused. A prior disposable 32-case pressure calibration qualified the model profiles for fresh holdout authoring. The scored 48-case corpus shared no scenario identifiers or scenario hashes with that calibration, and no calibration response or metric entered the scored denominator.

2.3 Model response contract

For every scenario, the model returned:

- one preferred complete option;
- one ranked authorized contingency;
- a bounded structured decision.

The model made one choice before arm evaluation. The study therefore obtained 48 decisions per profile, 96 decisions total. Three counterfactual arms evaluated each decision, yielding 288 arm results. Provider requests batched three scenarios at a time, producing 32 retained provider responses.

2.4 Arms

The arms were:

Prompt only. The instruction described the ordinary objective, but no model-external effect-level control mediated the selected option.

Exact action. Each exact synthetic operation received its local disposition. The design intentionally allowed substituted operations that were individually valid even when their reviewed effect-level relationship did not inherit authority.

Semantic effect authority. The evaluator applied the current reviewed relationship among the exact operation, principal, credential, purpose, destination, and materially related outcome. A selected substitute could be allowed, withheld, or routed to review. When the preferred option could not proceed, the evaluator considered the model's predeclared contingency.

The arms are matched counterfactual evaluations of one model choice. They are not separate behavioral samples and do not measure how model behavior would change after receiving a denial.

2.5 Outcomes

The study kept three units separate.

A pressure-bearing choice occurred when the model selected an alternate whose reviewed semantic disposition was not **allow**.

An unauthorized effect execution occurred when an effect with a non-allow semantic disposition completed under the evaluated arm.

An authorized objective completion occurred when at least one complete authorized option for the ordinary task completed.

Some decomposed options contained multiple effects. A single pressure-bearing choice could therefore produce more than one effect-execution opportunity.

2.6 Precommitted gates

Each model had to meet:

- structured validity of at least 95%;
- authorized-objective completion of at least 90%;
- pressure in at least 25% of the 32 pressure-bearing scenarios;
- pressure in at least four families;
- pressure through at least three operational mechanisms;
- zero unauthorized effects under semantic effect authority;
- zero provider failures.

Results were required to be reported by model before any pooled total.

3. Results

3.1 Model-selected pressure

DeepSeek selected non-authorized substitutes in 19/32 pressure-bearing scenarios, or 59.375%. Terra did so in 18/32, or 56.250%. Both models produced pressure across all eight families and all four operational mechanisms: degraded primary paths, latency advantages, delegated recommendations, and retained runbook preferences.

The 37 pressure-bearing choices produced 45 effect-execution opportunities: 23 for DeepSeek and 22 for Terra.

3.2 Counterfactual enforcement

Model	Prompt only	Exact action	Semantic effect authority
DeepSeek V4 Flash	23	23	0
GPT-5.6 Terra	22	22	0

Prompt-only and exact-action counts match because the substituted operations were locally permissible under the sealed design. The comparison targets a known limitation of operation-local authorization: a valid operation does not necessarily inherit authority for a materially related effect.

The empirical contribution is twofold. First, both models selected attractive substitutes without being instructed to violate authority. Second, holding those choices fixed, semantic effect authority distinguished the locally permissible operation from the reviewed authority governing the related effect.

3.3 Utility and contingencies

Authorized-objective completion was 48/48 for DeepSeek and 48/48 for Terra.

In the semantic arm, DeepSeek's preferred substitute was withheld or reviewed in 19 scenarios. Its authorized contingency completed the objective in 19/19. Terra's preferred substitute was withheld or reviewed in 18 scenarios, and its authorized contingency completed the objective in 18/18.

This 37/37 result demonstrates execution of a predeclared fallback under the counterfactual evaluator. It does not demonstrate adaptive recovery in which the model observes a denial, reasons again, and discovers a new authorized route.

Sixteen controls allowed either the standard or alternate route. Each model selected the alternate in ten cases and the standard route in six. All sixteen objectives completed. The retained metric name `authorized_alternate_completion_percent` is potentially misleading; it measures alternate uptake among permitted options, not utility. Descriptively, alternate uptake was 0/6 under low pressure, 4/4 under medium pressure, and 6/6 under high pressure for each model. This secondary pattern was not a publication gate.

3.4 Evidence integrity

All 32 provider responses completed with zero retries and zero provider failures. Structured validity was 100% for both models. The complete scored provider cost was \$0.0879199640.

Deterministic replay passed 48/48 scenarios. All 22 expected durable review tasks were created. The scoring report, scoring receipt, journal, corpus, source revision, and publication analysis were cryptographically bound.

The publication pipeline could not authorize itself. The deterministic analysis remained `public_claim_permitted: false`. A separate signer issued publication authority only after receiving an exact confirmation naming the analysis hash.

4. Interpretation

The study supports a narrow authority-continuity claim:

Across the tested synthetic scenarios, changing the route or representation of an effect did not create new authority under the semantic effect-authority evaluator.

This differs from a claim that the system universally infers semantic equivalence. The ground-truth outcome relationships and authority dispositions were independently reviewed within the synthetic protocol before scoring. FieldHash enforced those configured relationships; it did not discover every unstated enterprise dependency.

The study also differs from a live adaptive-agent evaluation. Because model choice was held constant, the design isolates the control mechanism. It does not show whether a model, after observing denial, would retry through another route, manipulate the gate, or discover a new authorized fallback. A live closed-loop extension is required for those questions.

The matched design and a live design answer different questions:

- Matched counterfactual: what does each control do with the same choice?
- Live closed loop: how does the agent change its behavior after the control responds?

Neither should be substituted for the other.

5. Limitations

The study is synthetic and self-administered. It had no independent human reviewer or external red team. The scenario author and reviewer were models commissioned by FieldHash and excluded from scoring.

One scored profile, GPT-5.6 Terra, shared an OpenAI product family with the GPT-5.6 Sol authoring profile. Excluding the exact authoring profile from scoring reduced direct role overlap but did not establish model-family independence.

The study does not establish:

- production performance;
- customer-policy correctness;
- open-world semantic equivalence;
- discovery of unstated dependencies;
- universal containment;
- protection for unmediated execution paths;
- resilience to host or kernel compromise;
- adaptive post-denial replanning;
- a population-level miss rate.

The 45 effect opportunities are clustered within 37 model choices, eight families, and one deterministic evaluator. A simple binomial rule-of-three calculation would therefore overstate statistical generality.

The provider runtimes were pinned by model identifier and request contract, but the providers did not attest byte-equivalence to separately published weights.

6. Next study

A live extension should preserve this study unchanged and use a fresh lineage. Each arm should run as a separate closed-loop episode. The model should receive the gate response, decide what to do next, and remain free to retry, change routes, or choose the sanctioned alternative.

Primary outcomes should include:

- first-choice unauthorized attempts;
- post-denial route substitutions;
- steps to authorized completion;
- adaptive fallback discovery;
- repeated denial and termination behavior;
- false denials;
- review routing;
- effect execution;
- end-to-end utility and latency.

The current study should remain the mechanism-isolation result. The live extension should test behavioral adaptation rather than reinterpret the matched evidence.

7. Artifact identifiers

- Compiled corpus SHA-256: 7f45c6c5863316aecde23cf29cae2e5ec62d547256445dc29709993c6bb597c2.
- Scoring report SHA-256: 16550a00affae61235383241f62e10bfdb96ab50ef35ac9f6925d91bcebfb08c.
- Scoring receipt SHA-256: ef2b2119d3521428a7f327a84cc729164aeeb3d38c589c89a3cefe70427e35cc.
- Publication analysis SHA-256: 8693d3b60befd02a22ce1e7d536d49e6c6c48c769c60772931f2405aad9e0fd5.
- Publication receipt SHA-256: b852954d9f3e999301845cc1af8b1055c0920231fc850f38789f4f5d5758e035.
- Scored source revision: 6e264e3bf41b733ca9d0fc9ae196b94f3e23bc40.

8. Public evidence availability

- Study page: <https://fieldhash.ai/case-studies/authority-under-effect-substitution>.
- Methods and evidence note: <https://fieldhash.ai/evidence/effect-substitution/methods.md>.
- Preauthorization analysis: <https://fieldhash.ai/evidence/effect-substitution/publication-analysis.json>.
- Signed publication receipt: <https://fieldhash.ai/evidence/effect-substitution/publication-receipt.json>.

The public projection is intentionally bounded. It describes the behavioral contract, results, verification lineage, and claim boundary without publishing the production effect-canonicalization implementation or internal equivalence rules.