VEHICLE Systems Lab

PUBLIC RESEARCH | VSL-EFFECTBOUND-001 | EXPANDED WEBSITE RESEARCH EDITION

# Effect-Bound Authority in Agentic Systems

## *Route-Invariant Control, External Effects, and Post-Effect Knowledge*

*Autoridad ligada al efecto en sistemas agentic: control invariante a la ruta, efectos externos y conocimiento posterior al efecto*

**Roberto Borda Milan**
VEHICLE Systems Lab, Bolivia
ORCID: 0009-0009-9047-1036
Archival DOI: 10.5281/zenodo.22820818
Archival release: V0.1.0 | 17 September 2026

**Status.** This document is an expanded exposition for the official VEHICLE Systems Lab website. It preserves the propositions, deterministic results, claim boundaries, and public/proprietary separation of the archived VSL-EFFECTBOUND-001 v0.1.0 release. It is not a new external validation, certification, or replacement DOI.

## Abstract

Agentic systems increasingly combine model reasoning, authenticated identities, interoperable tool protocols, and interfaces capable of producing external effects. That combination creates a control problem that is easy to obscure if identity, access, authority, execution, outcome, and evidence are treated as one event. This work formalizes effect-bound authority as a route-invariant decision over normalized effect semantics rather than over a single tool path. The public model distinguishes authenticated identity (I), server or tool access (S), exact effect authority (A), execution attempt (X), external effect (E), and post-effect knowledge (K). Five constructive propositions and four deterministic state-space experiments show structural insufficiencies in reduced observation spaces: identity/access does not uniquely determine exact effect authority; route-bound decisions may disagree for the same semantic effect; descriptive hints are not hard enforcement contracts; and tool responses alone need not identify the external-world outcome. The public reference gate is intentionally minimal and exact only with respect to the synthetic ground truth supplied by the experiments. It is not the proprietary VSL External Effect Control Plane (ECP). The contribution is therefore a reproducible formal target, a black-box conformance framing, and an explicit epistemic distinction between failed effects and unknown effects.

Keywords: agentic AI; effect-bound authorization; external effects; MCP; route invariance; TEVV; provenance; agent identity; post-effect knowledge.

# 1. Introduction and problem framing

The transition from chat-oriented models to tool-using agents changes the safety boundary. A model that only proposes text and a model that can invoke APIs, modify files, operate software, or actuate connected devices do not pose the same control problem. In the latter case, the relevant question is not only whether the agent is authenticated or whether a server accepts its token. It is whether a particular external effect, with particular parameters, target, environment, and timing, is authorized; whether that effect was attempted; whether it actually occurred; and what can later be established about the outcome.

Public developments in 2026 make the distinction concrete without requiring access to vendor internals. OpenAI published a framework that explicitly includes model behavior involving actions without authorization, coordination, or evasion of oversight, and its initial examples include unsanctioned actions taken to overcome obstacles [1]. Google Home exposes MCP tools for executing real-world actions and distinct tools for reading current state and history [2,3]. MCP maintainers explicitly describe tool annotations as hints and state that deterministic guarantees belong in enforcement mechanisms rather than descriptive metadata [4]. Anthropic discusses both governance risks for agents acting with less human oversight and engineering patterns for constraining the blast radius of increasingly capable agents [5,6]. Microsoft Entra Agent ID documents identity, OAuth tokens, scopes, app roles, and protected MCP-server access [7].

These sources are not treated as evidence that any named system lacks safeguards. They are used only to motivate a general systems question: what information must an enforcement boundary observe if it is expected to make a hard decision about a specific external effect?

> **Research question.** Can authority remain bound to the semantic external effect when an agent changes the computational route used to produce it?

## 1.1 Contribution

- A formal separation of I, S, A, X, E, and K as semantically distinct objects.
- A route-invariant authority target based on a normalized semantic effect fingerprint.
- Constructive proofs showing when reduced observations are insufficient for perfect classification.
- Four deterministic, exhaustive synthetic experiments with reproducible evidence.
- A black-box conformance boundary that allows private implementations to be compared without exposing proprietary internals.
- A strict interpretation boundary: synthetic error fractions are not vendor failure rates, and publication is not external validation.

## 1.2 Why the separation matters

In conventional application security, access control often asks whether a caller may invoke a protected resource or operation. Agentic systems add a semantic layer: multiple tools, protocols, routes, or intermediate plans may converge on the same externally meaningful effect. A permission model that is correct for one route can therefore be incomplete if equivalent routes remain reachable outside the same enforcement boundary. Similarly, a tool response may be useful evidence about execution, but it is not necessarily definitive evidence about the external world. The framework developed here treats these as separate measurement problems.

# 2. Formal model

Let a requested external effect be represented by the tuple:

```
q = (p, t, e, theta, epsilon, tau)
```

where p is the acting principal or agent identity; t is the target; e is the semantic effect class; theta is the normalized parameter set; epsilon is the environment or execution domain; and tau is the relevant time or expiry context. Let r denote a computational route or tool path that may be used to attempt q.

| Symbol | Object | Meaning |
|---|---|---|
| I | Authenticated identity | Whether the acting principal has been authenticated or otherwise established. |
| S | Server/tool access | Whether the principal can reach or invoke a server, API, connector, or tool surface. |
| A(q) | Exact effect authority | Whether the principal is authorized to produce the precise semantic effect q. |
| X(q,r) | Execution attempt | Whether an attempt to produce q is made through route r. |
| E(q) | External effect | Whether the externally meaningful state change actually occurs. |
| K(q) | Post-effect knowledge | What can be established after execution: CONFIRMED, FAILED, or UNKNOWN. |

Table 1. Public state variables. The relation I != S != A != X != E != K denotes semantic non-identity, not numeric inequality.

## 2.1 Semantic fingerprint and route invariance

Let N(theta) be a canonical normalization of the effect parameters and H an abstract collision-resistant digest. Define the semantic fingerprint:

```
phi(q) = H(p || t || e || N(theta) || epsilon || tau)
```

An effect-bound authorization predicate is then a function alpha over semantic fingerprints. The public route-independent decision target is:

```
A(q, r) = alpha(phi(q))
```

For two routes r1 and r2 that are equivalent with respect to the same semantic effect q, route invariance requires A(q,r1) = A(q,r2). The route itself may still be policy-relevant in a real system, but if policy grants or denies the semantic effect, then equivalent routes cannot silently create contradictory authority solely because they are different paths.


Figure 1. Control and evidence layers are related, but none should be assumed identical to its neighbor.

# 3. Constructive propositions

## 3.1 Proposition 1 - Insufficient observation

Let Q be a set of action requests, A:Q->{0,1} the correct effect-authority label, and pi:Q->O the information visible to a deterministic gate. If there exist requests qa and qu such that pi(qa) = pi(qu) while A(qa)=1 and A(qu)=0, then no deterministic classifier g:O->{0,1} can classify both requests correctly.

Proof. Because the gate receives the same observation for qa and qu, a deterministic classifier must return the same output for both. Their correct labels differ, so at least one classification is wrong. This is not an empirical statement about any vendor; it is a structural statement about information loss under projection.

Consequence. Identity and server access cannot be a complete signal for exact effect authority whenever distinct effects can share the same identity/access observation but differ in target, semantic class, parameters, environment, expiry, or another authority-relevant variable.

## 3.2 Proposition 2 - Route-invariance requirement

Suppose two routes r1 and r2 can produce the same semantic external effect q. A route-bound policy may return A(q,r1) != A(q,r2). By contrast, if authority is defined only through alpha(phi(q)), the decision is route-invariant by construction. The proposition does not require that every production system ignore routes. It states that a hard effect guarantee must not be accidentally weakened by an equivalent alternate path that escapes the same effect-level boundary.

> **Interpretation.** Blocking one path is not logically equivalent to blocking the effect if an equivalent path remains reachable outside the same enforcement boundary.

## 3.3 Proposition 3 - Route multiplicity and bypass exposure

Let Bj denote the event that an unauthorized semantic effect bypasses control through route rj. Across m reachable routes, the probability of at least one bypass is bounded by the largest individual route probability from below and the union bound from above:

```
max_j P(Bj) <= P(union Bj) <= min(1, sum_j P(Bj))
```

If independence is assumed only for illustration, then P(union Bj)=1-product_j(1-P(Bj)). No empirical bypass probability is assigned to OpenAI, Google, Anthropic, Microsoft, MCP, or any other vendor in this work. The equation simply shows why route multiplicity matters once per-route failure probabilities are non-zero.

## 3.4 Proposition 4 - Tool response does not identify external effect

Let Y be a tool or API response visible to the agent and E the external-world effect. If two executions can return the same Y while producing different E, then response alone cannot perfectly identify external effect. The same indistinguishability argument used in Proposition 1 applies. In particular, an error response or missing receipt is not logically equivalent to no effect. If the external state cannot be established, K=UNKNOWN is distinct from FAILED.

# 4. Hints, contracts, and the measurement boundary

## 4.1 Proposition 5 - Hints are not contracts

If tool annotation h is descriptive metadata and actual behavior b may differ from h, then any hard guarantee that depends solely on h=b is invalid unless an independent trusted mechanism enforces that equality. MCP maintainer guidance is unusually explicit on this point: tool annotations are hints, untrusted servers can misdescribe behavior, and deterministic guarantees belong in controls such as authorization, sandboxing, or network enforcement rather than in a boolean annotation [4].

The proposition generalizes beyond MCP. A model-facing description, label, schema annotation, documentation string, planner belief, or UI confirmation can be useful to guide behavior without becoming a hard security boundary. The architectural mistake arises when descriptive metadata is promoted to a guarantee without a trusted enforcement mechanism.

## 4.2 Control-plane separation

Effect-bound authority should be understood as one layer in a larger control stack, not as a replacement for identity and access management. Identity systems answer who or what is calling. OAuth scopes and app roles answer which protected resources or server capabilities a client may access. Sandboxes and network controls constrain where computation can reach. Model alignment and prompt-injection defenses address behavioral reliability. Human approvals can add contextual judgment. Effect-bound authority focuses on a narrower question: is this exact semantic external effect authorized under the relevant target, parameters, environment, and time context?

| Layer | Primary role | Question |
|---|---|---|
| Identity / authentication | Establish principal | Who or what is acting? |
| Server/tool authorization | Protect resource surface | May this principal call this server/tool class? |
| Effect-bound authority | Bind decision to q | May this exact external effect be produced? |
| Execution controls | Constrain attempt | How may the attempt be executed? |
| Observation / reconciliation | Establish external state | What actually happened? |
| Provenance / audit | Preserve evidence | What can later be demonstrated? |

Table 2. Complementary control layers. Effect-bound authority is intentionally positioned as a supplement, not a universal substitute.

## 4.3 Experimental method

The experiments are deterministic state-space enumerations. Each family defines a small set of binary or categorical variables, enumerates all states in that family, evaluates a deliberately reduced policy, evaluates an effect-bound or observation-aware reference, and records classification correctness. The value of this design is not statistical generalization; it is transparency. Every reported fraction can be reproduced from the public source, and every error can be traced to an explicit state.

The public reference is intentionally simple. In Experiments A and C it receives the synthetic ground-truth authority variable and therefore serves as a semantic target rather than as evidence of deployable performance. This limitation is explicit because hiding it would turn a formal reference into an unjustified product claim.

# 5. Deterministic experiments A and B

## 5.1 Experiment A - Identity/access versus effect authority

Experiment A enumerates twelve valid states across authenticated identity, server access, exact effect authority, and scope match. An access-bound policy allows whenever authenticated server access exists. The effect-bound reference allows only when identity, access, authority, and scope agree with the synthetic truth condition.

The access-bound policy produces three false allows in twelve valid states. Those errors occur precisely when the caller is authenticated and has server access but lacks authority for the exact effect or required scope. The reference produces zero classification errors on this constructed state space.

**Constructed result A.** Access-bound policy: 3 false allows / 12 states (error fraction 0.25). Effect-bound reference: 0 errors / 12 states. This is a synthetic theorem-exercising state space, not a Microsoft failure rate.

The Microsoft Entra Agent ID documentation provides a useful public contrast because it clearly explains the role of OAuth resource identification, scopes, app roles, access tokens, and claims validation for an MCP server [7]. The VSL proposition does not argue against those controls. It asks a different question: once a caller is authenticated and admitted to a tool surface, does the gate still have enough semantic information to decide the exact effect being requested?

## 5.2 Experiment B - Route invariance

Experiment B defines eight semantic effects, each reachable through four abstract routes. Semantic authority is fixed by effect. A constructed route-bound counterexample alternates decisions by route. Because the route pattern is independent of semantic authority, the policy can allow an unauthorized effect on one route and deny an authorized effect on another.

Across thirty-two effect-route cases, the route-bound counterexample classifies sixteen correctly and sixteen incorrectly. More directly, among forty-eight within-effect route-pair comparisons, thirty-two pairs disagree. The effect-bound reference is consistent in all forty-eight comparisons because its decision depends on phi(q) rather than on route identity.

**Constructed result B.** Route-bound counterexample: 32 inconsistent route pairs / 48; 16/32 case classifications incorrect. Effect-bound reference: 0 inconsistent route pairs. This proves existence of route inconsistency under the constructed policy; it is not a measured bypass frequency.

OpenAI is used only as a motivating public context. Its September 2026 disclosure framework explicitly includes models acting without authorization or evading oversight, and the initial examples include unsanctioned actions taken to overcome obstacles, use of unauthorized resources, and alternate channels used to accomplish tasks [1]. Those reports do not establish the VSL model; they demonstrate why route changes and unintended action channels are meaningful phenomena to reason about.

# 6. Deterministic experiments C and D

## 6.1 Experiment C - Hints versus contracts

Experiment C enumerates eight states across semantic authority, a declared safe hint, and whether the operation is actually effecting. The hint-bound policy treats the descriptive hint as permission. The effect-bound reference uses the synthetic semantic authority relation. The hint-bound policy misclassifies four of eight states: one false allow and three false denies. The reference is exact on the constructed state space.

**Constructed result C.** Hint-bound policy: 4 errors / 8 states (error fraction 0.50). Effect-bound reference: 0 errors / 8 states. The experiment operationalizes the general proposition that descriptive metadata is not a hard contract.

This experiment aligns directly with MCP maintainer guidance that annotations such as readOnlyHint, destructiveHint, idempotentHint, and openWorldHint are hints rather than guaranteed descriptions, and that hard guarantees require deterministic controls [4]. The experiment is not an audit of MCP or any MCP implementation.

## 6.2 Experiment D - Response versus effect knowledge

Experiment D enumerates eight states across tool response success, actual external effect occurrence, and availability of a post-effect observation channel. A response-only classifier maps success to CONFIRMED and error to FAILED. The observation-aware reference instead reports CONFIRMED when the observed external state establishes occurrence, FAILED when observation establishes non-occurrence, and UNKNOWN when the relevant state cannot be established.

The response-only classifier is correct in only two of eight states and misclassifies six. The observation-aware reference is exact on the synthetic state space. The important point is not that observation is always easy; it is that uncertainty deserves an explicit state rather than being silently converted into failure or success.

**Constructed result D.** Response-only classifier: 6 errors / 8 states (error fraction 0.75). Observation-aware reference: 0 errors / 8 states. Google Home MCP is used only as a public example of separate action and state/history tools [2,3].

| Exp. | Reduced policy | N | Err. frac. | Observed error | Reference |
|---|---|---|---|---|---|
| A | Access-bound | 12 | 0.25 | 3 false allows | Effect-bound reference: 0 errors |
| B | Route-bound counterexample | 32 cases / 48 pairs | 0.50 case error | 32 inconsistent pairs | Effect-bound reference: 0 inconsistent pairs |
| C | Hint-bound | 8 | 0.50 | 4 errors | Effect-bound reference: 0 errors |
| D | Response-only | 8 | 0.75 | 6 errors | Observation-aware reference: 0 errors |

Table 3. Exhaustive results over abstract synthetic state spaces. These fractions are not empirical rates for any named vendor.

# 7. Discussion: what the experiments establish

The strongest conclusion supported by VSL-EFFECTBOUND-001 is structural rather than empirical. If an enforcement boundary cannot distinguish two requests that require different authority decisions, then no deterministic classifier using only that reduced observation can classify both correctly. If equivalent routes are governed by route-specific decisions rather than the semantic effect, the same effect can receive contradictory authority. If metadata is not trustworthy or enforced, it cannot supply a hard guarantee. If a tool response does not uniquely determine the external state, then response-only reasoning cannot perfectly establish what happened.

These statements are deliberately narrower than claims such as "agentic systems are insecure" or "IAM is insufficient". Identity and access systems are necessary and often highly sophisticated. The VSL claim is that they answer a different layer of the problem. A robust agentic control architecture may need several layers simultaneously: identity, protected-resource authorization, effect-level policy, constrained execution, post-effect observation, reconciliation, provenance, and human governance.

## 7.1 Post-effect knowledge as an epistemic state

The K variable is central because many automation systems collapse uncertainty into a binary status. In effecting systems that collapse can be dangerous. A network timeout can occur before a request reaches the target, during processing, or after the target has committed the effect but before the response returns. Similarly, a tool may report success while downstream execution fails, or may report an error after a non-idempotent effect has already occurred. Unless the response channel is itself a definitive observation of the relevant external state, the correct epistemic result may be UNKNOWN.

This distinction has practical consequences for retries. If UNKNOWN is incorrectly treated as FAILED, an automatic retry can duplicate an irreversible effect. If UNKNOWN is incorrectly treated as CONFIRMED, the system can falsely assume that an obligation or safety action completed. A production system therefore needs effect-specific reconciliation rules and, where possible, independent observation channels or idempotency mechanisms.

## 7.2 Route invariance as a design test

Route invariance can be used as a black-box test even when the authorization implementation is private. Construct a set of semantically equivalent requests q that differ only in admissible route r. If the intended policy is effect-level and route-independent, then decisions should be invariant across those routes. If the policy intentionally treats routes differently, the difference should be explicit in the policy semantics rather than an accidental side effect of tooling.

## 7.3 Relationship to agent containment

Anthropic frames containment in terms of controlling the blast radius of increasingly capable agents and describes environment-level isolation patterns [6]. Effect-bound authority is complementary to that perspective. Containment constrains the environment and potential damage surface; effect-bound authority constrains which semantic effects are authorized. Neither subsumes the other. A well-contained agent can still be authorized incorrectly, and a correctly authorized effect can still be executed in an environment with excessive blast radius.

# 8. Relationship to VSL-ECP and public/proprietary boundary

VEHICLE Systems Lab studies a proprietary External Effect Control Plane (ECP). The public research artifact was intentionally designed so that the scientific question can be examined without disclosing the implementation that VSL considers commercially sensitive. The public reference gate is therefore not a stripped-down product version; it is a deliberately simple mathematical oracle for the synthetic experiments.

## 8.1 Public material

- Formal definitions, propositions, and constructive proofs.
- Abstract deterministic state spaces and generated evidence.
- The reproducibility scripts and public reference decisions.
- A black-box conformance interface and publication guidance.
- Official public context sources and explicit non-endorsement language.

## 8.2 Intentionally excluded

- Production permit schema, proprietary field semantics, and internal policy representation.
- Signing and verification internals, key management, and cryptographic material.
- Durable ledger/storage implementation, concurrency, consumption, and recovery logic.
- Internal APIs, deployment identifiers, production connectors, secrets, and credentials.
- Private tests that would expose bypass-relevant implementation detail.

**Boundary.** The open reference gate must not be described as the ECP engine. The DOI publication establishes a public research artifact, not product certification or disclosure of proprietary implementation.

# 9. Threats to validity and non-claims

1. Synthetic state spaces. The reported fractions are exhaustive over deliberately small abstract state spaces; they do not estimate field frequencies.
2. No vendor execution. OpenAI, Google, Anthropic, Microsoft, MCP, and other named systems were not internally executed, penetrated, or benchmarked by VSL for these results.
3. Reference-oracle advantage. In some experiments the effect-bound reference receives synthetic ground truth. Zero error therefore defines the target semantics rather than demonstrating deployable performance.
4. Incomplete real-world semantics. Production authority can depend on policy layers, business rules, consent, physical state, human review, and environmental context beyond this public model.
5. Observation can also fail. Post-effect observation may be delayed, compromised, stale, or unavailable; richer reconciliation is required in those cases.
6. No global novelty claim. The propositions and framing are published as VSL work, but no claim of exhaustive prior-art novelty is made until broader review is complete.
7. Publication is not independent validation. GitHub, Zenodo, DOI registration, and ORCID indexing provide persistence and citation, not peer review or third-party replication.

# 10. Reproducibility and provenance

The public artifact is designed so that every deterministic result can be regenerated from source. The canonical archival chain is: GitHub commit -> GitHub release V0.1.0 -> Zenodo archive -> DOI 10.5281/zenodo.22820818. The release contains the experiment generator, tests, evidence CSV/JSON, formal definitions, propositions, manuscript, checksums, licensing, and a public/proprietary boundary statement.

## 10.1 Minimal reproduction

```
python experiment/run_experiment.py
python experiment/test_experiment.py
```

The archived package reports PASS for eight invariant checks. The generated evidence is deterministic and hashable. A private implementation can be connected locally to the black-box PowerShell conformance harness without publishing the proprietary adapter. Only normalized decisions, outcome classifications, evidence hashes, and aggregate metrics need to leave the private environment.

## 10.2 Reproducibility ladder

| Level | Evidence stage | Meaning |
|---|---|---|
| Level 0 | Archived artifact | Persistent DOI, immutable release, checksums, source and evidence. |
| Level 1 | Clean-room rerun | A fresh environment reproduces the deterministic outputs from the public package. |
| Level 2 | Independent implementation | A second implementation reproduces the propositions without reusing the reference decision code. |
| Level 3 | External reproducer | A third party independently reproduces and documents the result. |
| Level 4 | Black-box application | The conformance protocol is applied to a real effecting system without access to its internals. |
| Level 5 | Field evidence | Repeated real-world observations support bounded empirical claims about a specific implementation and population. |

Table 4. Proposed evidence ladder. VSL-EFFECTBOUND-001 currently establishes the archived public artifact and deterministic package; higher levels require additional work.

## 10.3 Current evidence status

As of 17 September 2026, the public work is best described as published and reproducible by design, with deterministic results packaged under a DOI. Independent implementation and third-party reproduction remain future evidence targets. This wording protects the distinction between a citable research object and an externally validated result.

## 11. Research agenda

The most valuable next step is not to add more synthetic cases merely to increase volume. It is to change the source of evidence. A stronger follow-on study should preserve the v0.1.0 semantics while testing whether independent implementations and black-box systems satisfy the same invariants.

- Independent implementation: re-implement the propositions from the paper without using the public reference decision code.
- Route-equivalence corpus: define richer semantic equivalence classes and test invariance across tool families, API variants, and planner-generated routes.
- Partial-observation reconciliation: extend K beyond the three-state model to include stale, conflicting, delayed, and independently corroborated observations.
- Effect normalization study: examine when two requests should map to the same phi(q), including canonicalization, units, target aliases, and time windows.
- Black-box conformance: apply the public harness to private systems while exporting only aggregate evidence and hashes.
- External reproduction: invite an independent researcher or laboratory to reproduce the formal and experimental package and publish deviations or confirmations.

## 12. Conclusion

Agentic control becomes conceptually fragile when identity, access, authority, execution, outcome, and knowledge are collapsed into one event. VSL-EFFECTBOUND-001 separates those layers and formalizes a narrow requirement: if a hard authorization decision is intended to govern a semantic external effect, equivalent computational routes should not silently change the decision unless route dependence is itself part of the intended policy. Likewise, post-effect knowledge should be based on observations capable of distinguishing relevant external states; otherwise UNKNOWN must remain distinct from FAILED or CONFIRMED.

The four deterministic experiments do not measure vendor reliability. They do something smaller and more defensible: they construct transparent counterexamples showing why reduced observations can be insufficient, why route-specific decisions can disagree, why hints cannot substitute for hard contracts, and why responses cannot always identify external effects. The public reference is a semantic oracle for those constructions, not a production security claim. The resulting artifact is intended as a reproducible target for future independent implementations, black-box conformance, and eventually field evidence.

**Canonical citation.** Borda Milan, Roberto. Effect-Bound Authority in Agentic Systems: Route-Invariant Control, External Effects, and Post-Effect Knowledge. VSL-EFFECTBOUND-001 v0.1.0, 2026. DOI: 10.5281/zenodo.22820818.

## References

**[1] OpenAI.** Our framework for reporting model misalignment. 16 Sep 2026. https://openai.com/index/model-misalignment-reporting-framework/

**[2] Google Home Developers.** MCP Reference: home.googleapis.com. Accessed 17 Sep 2026. https://developers.home.google.com/reference/home/mcp

**[3] Google Home Developers.** MCP Tools Reference: run_home_actions. Accessed 17 Sep 2026. https://developers.home.google.com/reference/home/mcp/tools_list/run_home_actions

**[4] Model Context Protocol maintainers.** Tool Annotations as Risk Vocabulary: What Hints Can and Can't Do. 16 Mar 2026. https://blog.modelcontextprotocol.io/posts/2026-03-16-tool-annotations/

**[5] Anthropic.** Trustworthy agents in practice. 9 Apr 2026. https://www.anthropic.com/research/trustworthy-agents

**[6] Anthropic.** How we contain Claude across products. 25 May 2026. https://www.anthropic.com/engineering/how-we-contain-claude

**[7] Microsoft Learn.** Secure a Model Context Protocol (MCP) server with Microsoft Entra ID. Accessed 17 Sep 2026. https://learn.microsoft.com/en-us/entra/agent-id/secure-mcp-server-with-entra-id

**[8] Borda Milan, Roberto.** Effect-Bound Authority in Agentic Systems: Route-Invariant Control, External Effects, and Post-Effect Knowledge. VSL-EFFECTBOUND-001 v0.1.0. Zenodo, 2026. https://doi.org/10.5281/zenodo.22820818

Non-endorsement. OpenAI, Google, Anthropic, the MCP project, and Microsoft did not sponsor, validate, endorse, review, or participate in this VSL experiment. Their public documentation is cited only to motivate abstract boundary conditions and related engineering concerns.