Autoridad ligada al efecto en sistemas agentic

Control invariante a la ruta, efectos externos y conocimiento posterior al efecto

Original archivado en inglés: Effect-Bound Authority in Agentic Systems: Route-Invariant Control, External Effects, and Post-Effect Knowledge

Roberto Borda Milan
VEHICLE Systems Lab, Bolivia
ORCID: 0009-0009-9047-1036
DOI archivado: 10.5281/zenodo.22820818
Release archivado: V0.1.0 | 17 de septiembre de 2026

Estado. Esta es la edición espejo en español para el sitio oficial de VEHICLE Systems Lab. Reproduce la misma estructura científica, proposiciones, resultados deterministas, límites de afirmación y separación pública/propietaria de la edición web ampliada en inglés. El DOI identifica el release archivado VSL-EFFECTBOUND-001 v0.1.0 en inglés; esta traducción no crea un DOI nuevo ni constituye una nueva validación, certificación o versión experimental.

Resumen

Los sistemas agentic combinan cada vez más razonamiento de modelos, identidades autenticadas, protocolos interoperables de herramientas e interfaces capaces de producir efectos externos. Esa combinación crea un problema de control que puede quedar oculto si identidad, acceso, autoridad, ejecución, resultado y evidencia se tratan como un solo evento. Este trabajo formaliza la autoridad ligada al efecto como una decisión invariante a la ruta sobre semántica normalizada del efecto, en lugar de vincularla a una única ruta de herramienta. El modelo público distingue identidad autenticada (I), acceso a servidor o herramienta (S), autoridad sobre el efecto exacto (A), intento de ejecución (X), efecto externo (E) y conocimiento posterior al efecto (K). Cinco proposiciones constructivas y cuatro experimentos deterministas sobre espacios de estados muestran insuficiencias estructurales en espacios de observación reducidos: identidad/acceso no determina de forma única la autoridad exacta del efecto; decisiones ligadas a la ruta pueden discrepar para el mismo efecto semántico; los hints descriptivos no son contratos duros de enforcement; y las respuestas de herramientas, por sí solas, no necesariamente identifican el resultado en el mundo externo. La compuerta de referencia pública es deliberadamente mínima y exacta solo respecto de la verdad sintética suministrada por los experimentos. No es el External Effect Control Plane (ECP) propietario de VSL. La contribución es, por tanto, un objetivo formal reproducible, un marco de conformidad de caja negra y una distinción epistémica explícita entre efectos fallidos y efectos cuyo estado es desconocido.

Palabras clave: IA agentic; autorización ligada al efecto; efectos externos; MCP; invariancia de ruta; TEVV; procedencia; identidad de agente; conocimiento posterior al efecto.

1. Introducción y planteamiento del problema

La transición desde modelos orientados al chat hacia agentes que utilizan herramientas cambia la frontera de seguridad. Un modelo que solo propone texto y un modelo capaz de invocar API, modificar archivos, operar software o accionar dispositivos conectados no presentan el mismo problema de control. En el segundo caso, la pregunta relevante no es solo si el agente está autenticado o si un servidor acepta su token. También importa si un efecto externo concreto, con determinados parámetros, objetivo, entorno y momento, está autorizado; si ese efecto fue intentado; si realmente ocurrió; y qué puede establecerse posteriormente sobre el resultado.

Los desarrollos públicos de 2026 permiten concretar esta distinción sin requerir acceso a los sistemas internos de los proveedores. OpenAI publicó un marco que incluye explícitamente comportamientos de modelos relacionados con acciones sin autorización, coordinación o evasión de supervisión, y sus ejemplos iniciales incluyen acciones no autorizadas para superar obstáculos [1]. Google Home expone herramientas MCP para ejecutar acciones en el mundo real y herramientas distintas para leer estado actual e historial [2,3]. Los mantenedores de MCP describen explícitamente las annotations de herramientas como hints y señalan que las garantías deterministas corresponden a mecanismos de enforcement y no a metadata descriptiva [4]. Anthropic analiza tanto riesgos de gobernanza de agentes que operan con menor supervisión humana como patrones de ingeniería para limitar el radio de impacto de agentes cada vez más capaces [5,6]. Microsoft Entra Agent ID documenta identidad, tokens OAuth, scopes, app roles y acceso protegido a servidores MCP [7].

Estas fuentes no se tratan como evidencia de que algún sistema nombrado carezca de salvaguardas. Se utilizan únicamente para motivar una pregunta general de sistemas: ¿qué información debe observar una frontera de enforcement si se espera que tome una decisión dura sobre un efecto externo específico?

> **Pregunta de investigación.** ¿Puede la autoridad permanecer ligada al efecto externo semántico cuando el agente cambia la ruta computacional utilizada para producirlo?

1.1 Contribución

- Una separación formal de I, S, A, X, E y K como objetos semánticamente distintos.
- Un objetivo de autoridad invariante a la ruta basado en una huella semántica normalizada del efecto.
- Pruebas constructivas que muestran cuándo las observaciones reducidas son insuficientes para una clasificación perfecta.
- Cuatro experimentos sintéticos deterministas y exhaustivos con evidencia reproducible.
- Una frontera de conformidad de caja negra que permite comparar implementaciones privadas sin exponer sus componentes internos propietarios.
- Una frontera estricta de interpretación: las fracciones de error sintéticas no son tasas de fallo de proveedores y la publicación no constituye validación externa.

1.2 Por qué importa la separación

En seguridad de aplicaciones convencional, el control de acceso suele preguntar si un llamador puede invocar un recurso u operación protegidos. Los sistemas agentic añaden una capa semántica: múltiples herramientas, protocolos, rutas o planes intermedios pueden converger en el mismo efecto con significado externo. Un modelo de permisos correcto para una ruta puede ser incompleto si rutas equivalentes continúan siendo alcanzables fuera de la misma frontera de enforcement. Del mismo modo, una respuesta de herramienta puede aportar evidencia útil sobre la ejecución, pero no necesariamente constituye evidencia definitiva sobre el mundo externo. El marco desarrollado aquí trata estos puntos como problemas de medición separados.

2. Modelo formal

Sea un efecto externo solicitado representado por la tupla:

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

donde p es el principal actuante o identidad del agente; t es el objetivo; e es la clase semántica del efecto; theta es el conjunto normalizado de parámetros; epsilon es el entorno o dominio de ejecución; y tau es el contexto temporal o de expiración relevante. Sea r una ruta computacional o ruta de herramienta que puede utilizarse para intentar q.

Símbolo	Objeto	Significado
I	Identidad autenticada	Si el principal actuante ha sido autenticado o establecido de otro modo.
S	Acceso a servidor/herramienta	Si el principal puede alcanzar o invocar un servidor, API, conector o superficie de herramientas.
A(q)	Autoridad exacta del efecto	Si el principal está autorizado para producir el efecto semántico preciso q.
X(q,r)	Intento de ejecución	Si se realiza un intento de producir q mediante la ruta r.
E(q)	Efecto externo	Si el cambio de estado con significado externo ocurre realmente.
K(q)	Conocimiento posterior al efecto	Qué puede establecerse después de la ejecución: CONFIRMED, FAILED o UNKNOWN.

Tabla 1. Variables de estado públicas. La relación I != S != A != X != E != K expresa no-identidad semántica, no desigualdad numérica.

2.1 Huella semántica e invariancia de ruta

Sea N(theta) una normalización canónica de los parámetros del efecto y H un resumen abstracto resistente a colisiones. Definimos la huella semántica:

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

Un predicado de autorización ligada al efecto es entonces una función alpha sobre huellas semánticas. El objetivo público de decisión independiente de la ruta es:

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

Para dos rutas r1 y r2 equivalentes respecto del mismo efecto semántico q, la invariancia de ruta exige A(q,r1) = A(q,r2). La ruta puede seguir siendo relevante para la política en un sistema real, pero si la política concede o deniega el efecto semántico, rutas equivalentes no deberían crear silenciosamente autoridades contradictorias solo por ser caminos distintos.

Figura 1. Las capas de control y evidencia están relacionadas, pero no debe asumirse que alguna sea idéntica a la contigua.

3. Proposiciones constructivas

3.1 Proposición 1 - Observación insuficiente

Sea Q un conjunto de solicitudes de acción, A:Q->{0,1} la etiqueta correcta de autoridad sobre el efecto y pi:Q->O la información visible para una compuerta determinista. Si existen solicitudes qa y qu tales que pi(qa) = pi(qu), mientras A(qa)=1 y A(qu)=0, entonces ningún clasificador determinista g:O->{0,1} puede clasificar correctamente ambas solicitudes.

Prueba. Como la compuerta recibe la misma observación para qa y qu, un clasificador determinista debe devolver la misma salida para ambas. Sus etiquetas correctas difieren, por lo que al menos una clasificación es errónea. Esto no es una afirmación empírica sobre ningún proveedor; es una afirmación estructural sobre pérdida de información bajo proyección.

Consecuencia. Identidad y acceso al servidor no pueden constituir una señal completa de autoridad exacta sobre el efecto cuando efectos distintos pueden compartir la misma observación identidad/acceso pero diferir en objetivo, clase semántica, parámetros, entorno, expiración u otra variable relevante para la autoridad.

3.2 Proposición 2 - Requisito de invariancia de ruta

Supongamos que dos rutas r1 y r2 pueden producir el mismo efecto externo semántico q. Una política ligada a la ruta puede devolver A(q,r1) != A(q,r2). En cambio, si la autoridad se define solo mediante alpha(phi(q)), la decisión es invariante a la ruta por construcción. La proposición no exige que todo sistema de producción ignore las rutas. Establece que una garantía dura sobre el efecto no debe debilitarse accidentalmente por una ruta alternativa equivalente que escape de la misma frontera a nivel de efecto.

> **Interpretación.** Bloquear una ruta no equivale lógicamente a bloquear el efecto si una ruta equivalente sigue siendo alcanzable fuera de la misma frontera de enforcement.

3.3 Proposición 3 - Multiplicidad de rutas y exposición a elusión

Sea Bj el evento en que un efecto semántico no autorizado elude el control mediante la ruta rj. Entre m rutas alcanzables, la probabilidad de al menos una elusión está acotada inferiormente por la mayor probabilidad individual de ruta y superiormente por la cota de la unión:

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

Si la independencia se asume solo con fines ilustrativos, entonces P(union Bj)=1-product_j(1-P(Bj)). En este trabajo no se asigna una probabilidad empírica de elusión a OpenAI, Google, Anthropic, Microsoft, MCP ni a ningún otro proveedor. La ecuación solo muestra por qué la multiplicidad de rutas importa una vez que las probabilidades de fallo por ruta son distintas de cero.

3.4 Proposición 4 - La respuesta de herramienta no identifica el efecto externo

Sea Y una respuesta de herramienta o API visible para el agente y E el efecto en el mundo externo. Si dos ejecuciones pueden devolver el mismo Y y producir E diferentes, la respuesta por sí sola no puede identificar perfectamente el efecto externo. Se aplica el mismo argumento de indistinguibilidad de la Proposición 1. En particular, una respuesta de error o la ausencia de un recibo no equivale lógicamente a ausencia de efecto. Si el estado externo no puede establecerse, K=UNKNOWN es distinto de FAILED.

4. Hints, contratos y frontera de medición

4.1 Proposición 5 - Los hints no son contratos

Si una annotation h de herramienta es metadata descriptiva y el comportamiento real b puede diferir de h, entonces cualquier garantía dura que dependa solo de h=b es inválida salvo que un mecanismo confiable e independiente imponga esa igualdad. La guía de los mantenedores de MCP es especialmente explícita: las annotations de herramientas son hints, los servidores no confiables pueden describir incorrectamente su comportamiento y las garantías deterministas corresponden a controles como autorización, sandboxing o enforcement de red, no a una annotation booleana [4].

La proposición se generaliza más allá de MCP. Una descripción orientada al modelo, una etiqueta, una annotation de esquema, una cadena de documentación, una creencia del planificador o una confirmación de interfaz pueden ser útiles para guiar comportamiento sin convertirse en una frontera dura de seguridad. El error arquitectónico aparece cuando metadata descriptiva se eleva a garantía sin un mecanismo de enforcement confiable.

4.2 Separación del plano de control

La autoridad ligada al efecto debe entenderse como una capa dentro de una pila de control mayor, no como sustituto de la gestión de identidad y acceso. Los sistemas de identidad responden quién o qué realiza la llamada. Los scopes OAuth y app roles responden a qué recursos protegidos o capacidades de servidor puede acceder un cliente. Los sandboxes y controles de red limitan hasta dónde puede llegar la computación. El alineamiento del modelo y las defensas contra prompt injection abordan confiabilidad conductual. Las aprobaciones humanas pueden aportar juicio contextual. La autoridad ligada al efecto se centra en una pregunta más estrecha: ¿está autorizado este efecto externo semántico exacto bajo el objetivo, parámetros, entorno y contexto temporal relevantes?

Capa	Rol principal	Pregunta
Identidad / autenticación	Establecer principal	¿Quién o qué está actuando?
Autorización de servidor/herramienta	Proteger superficie de recursos	¿Puede este principal invocar esta clase de servidor/herramienta?
Autoridad ligada al efecto	Vincular decisión a q	¿Puede producirse este efecto externo exacto?
Controles de ejecución	Restringir el intento	¿Cómo puede ejecutarse el intento?
Observación / reconciliación	Establecer estado externo	¿Qué ocurrió realmente?
Procedencia / auditoría	Preservar evidencia	¿Qué puede demostrarse posteriormente?

Tabla 2. Capas de control complementarias. La autoridad ligada al efecto se posiciona intencionalmente como complemento, no como sustituto universal.

4.3 Método experimental

Los experimentos son enumeraciones deterministas de espacios de estados. Cada familia define un conjunto pequeño de variables binarias o categóricas, enumera todos los estados de esa familia, evalúa una política deliberadamente reducida, evalúa una referencia ligada al efecto o basada en observación y registra la corrección de la clasificación. El valor de este diseño no es la generalización estadística, sino la transparencia. Cada fracción reportada puede reproducirse desde la fuente pública y cada error puede rastrearse hasta un estado explícito.

La referencia pública es intencionalmente simple. En los Experimentos A y C recibe la variable sintética de verdad de autoridad y, por tanto, funciona como objetivo semántico, no como evidencia de rendimiento desplegable. Esta limitación se declara de forma explícita porque ocultarla convertiría una referencia formal en una afirmación injustificada de producto.

5. Experimentos deterministas A y B

5.1 Experimento A - Identidad/acceso frente a autoridad del efecto

El Experimento A enumera doce estados válidos entre identidad autenticada, acceso al servidor, autoridad exacta sobre el efecto y coincidencia de scope. Una política ligada al acceso autoriza siempre que existan identidad autenticada y acceso al servidor. La referencia ligada al efecto autoriza solo cuando identidad, acceso, autoridad y scope concuerdan con la condición de verdad sintética.

La política ligada al acceso produce tres autorizaciones falsas en doce estados válidos. Esos errores aparecen precisamente cuando el llamador está autenticado y tiene acceso al servidor, pero carece de autoridad sobre el efecto exacto o del scope requerido. La referencia produce cero errores de clasificación en este espacio de estados construido.

> **Resultado construido A.** Política ligada al acceso: 3 autorizaciones falsas / 12 estados (fracción de error 0.25). Referencia ligada al efecto: 0 errores / 12 estados. Es un espacio sintético construido para ejercitar la proposición, no una tasa de fallo de Microsoft.

La documentación de Microsoft Entra Agent ID ofrece un contraste público útil al explicar identificación de recursos OAuth, scopes, app roles, access tokens y validación de claims para un servidor MCP [7]. VSL no cuestiona esos controles; pregunta si, una vez autenticado y admitido el llamador, la compuerta conserva suficiente información semántica para decidir el efecto exacto solicitado.

5.2 Experimento B - Invariancia de ruta

El Experimento B define ocho efectos semánticos, cada uno alcanzable mediante cuatro rutas abstractas. La autoridad semántica se fija por efecto. Un contraejemplo construido ligado a la ruta alterna decisiones según la ruta. Como el patrón de ruta es independiente de la autoridad semántica, la política puede autorizar un efecto no autorizado por una ruta y denegar un efecto autorizado por otra.

En treinta y dos casos efecto-ruta, el contraejemplo ligado a la ruta clasifica dieciséis correctamente y dieciséis incorrectamente. De forma más directa, entre cuarenta y ocho comparaciones de pares de rutas dentro del mismo efecto, treinta y dos pares discrepan. La referencia ligada al efecto es consistente en las cuarenta y ocho comparaciones porque su decisión depende de phi(q) y no de la identidad de la ruta.

> **Resultado construido B.** Contraejemplo ligado a la ruta: 32 pares de rutas inconsistentes / 48; 16/32 clasificaciones de casos incorrectas. Referencia ligada al efecto: 0 pares inconsistentes. Esto demuestra existencia de inconsistencia de ruta bajo la política construida; no es una frecuencia de bypass medida.

OpenAI se utiliza solo como contexto motivador. Su marco de septiembre de 2026 incluye modelos que actúan sin autorización o evaden supervisión, y ejemplos iniciales de acciones no autorizadas para superar obstáculos, uso de recursos no autorizados y canales alternativos [1]. Esos reportes no establecen el modelo de VSL; muestran por qué los cambios de ruta y los canales imprevistos son fenómenos relevantes.

6. Experimentos deterministas C y D

6.1 Experimento C - Hints frente a contratos

El Experimento C enumera ocho estados entre autoridad semántica, un hint declarado seguro y si la operación realmente produce un efecto. La política basada en hints trata el hint descriptivo como permiso. La referencia ligada al efecto utiliza la relación sintética de autoridad semántica. La política basada en hints clasifica incorrectamente cuatro de ocho estados: una autorización falsa y tres denegaciones falsas. La referencia es exacta en el espacio de estados construido.

Resultado construido C. Política basada en hints: 4 errores / 8 estados (fracción de error 0.50). Referencia ligada al efecto: 0 errores / 8 estados. El experimento operacionaliza la proposición general de que metadata descriptiva no equivale a un contrato duro.

Este experimento se alinea directamente con la guía de los mantenedores de MCP, según la cual annotations como readOnlyHint, destructiveHint, idempotentHint y openWorldHint son hints y no descripciones garantizadas, y las garantías duras requieren controles deterministas [4]. El experimento no es una auditoría de MCP ni de ninguna implementación MCP.

6.2 Experimento D - Respuesta frente a conocimiento del efecto

El Experimento D enumera ocho estados entre éxito de la respuesta de herramienta, ocurrencia real del efecto externo y disponibilidad de un canal de observación posterior al efecto. Un clasificador basado solo en la respuesta asigna éxito a CONFIRMED y error a FAILED. La referencia basada en observación informa CONFIRMED cuando el estado externo observado establece la ocurrencia, FAILED cuando la observación establece la no ocurrencia y UNKNOWN cuando el estado relevante no puede establecerse.

El clasificador basado solo en respuesta acierta solo en dos de ocho estados y falla en seis. La referencia basada en observación es exacta en el espacio sintético. El punto importante no es que observar siempre sea fácil; es que la incertidumbre merece un estado explícito en lugar de convertirse silenciosamente en fallo o éxito.

Resultado construido D. Clasificador basado solo en respuesta: 6 errores / 8 estados (fracción de error 0.75). Referencia basada en observación: 0 errores / 8 estados. Google Home MCP se utiliza solo como ejemplo público de herramientas separadas de acción y de estado/historial [2,3].

Exp.	Política reducida	N	Frac. error	Error observado	Referencia
A	Ligada al acceso	12	0.25	3 autorizaciones falsas	Ligada al efecto: 0 errores
B	Contraejemplo ligado a ruta	32 casos / 48 pares	0.50 error por caso	32 pares inconsistentes	Ligada al efecto: 0 pares inconsistentes
C	Basada en hints	8	0.50	4 errores	Ligada al efecto: 0 errores
D	Solo respuesta	8	0.75	6 errores	Basada en observación: 0 errores

Tabla 3. Resultados exhaustivos sobre espacios de estados sintéticos abstractos. Estas fracciones no son tasas empíricas de ningún proveedor nombrado.

7. Discusión: qué establecen los experimentos

La conclusión más fuerte respaldada por VSL-EFFECTBOUND-001 es estructural y no empírica. Si una frontera de enforcement no puede distinguir dos solicitudes que requieren decisiones de autoridad distintas, ningún clasificador determinista que use solo esa observación reducida puede clasificar correctamente ambas. Si rutas equivalentes se gobiernan mediante decisiones específicas de ruta en lugar del efecto semántico, el mismo efecto puede recibir autoridades contradictorias. Si la metadata no es confiable ni está sometida a enforcement, no puede proporcionar una garantía dura. Si la respuesta de una herramienta no determina de forma única el estado externo, el razonamiento basado solo en respuesta no puede establecer perfectamente qué ocurrió.

Estas afirmaciones son deliberadamente más estrechas que afirmaciones como "los sistemas agentic son inseguros" o "IAM es insuficiente". Los sistemas de identidad y acceso son necesarios y con frecuencia muy sofisticados. La afirmación de VSL es que responden a una capa distinta del problema. Una arquitectura robusta de control agentic puede necesitar simultáneamente varias capas: identidad, autorización de recurso protegido, política a nivel de efecto, ejecución restringida, observación posterior al efecto, reconciliación, procedencia y gobernanza humana.

7.1 Conocimiento posterior al efecto como estado epistémico

La variable K es central porque muchos sistemas de automatización colapsan la incertidumbre en un estado binario. En sistemas que producen efectos, ese colapso puede ser peligroso. Un timeout de red puede ocurrir antes de que una solicitud llegue al objetivo, durante el procesamiento o después de que el objetivo haya confirmado el efecto pero antes de que regrese la respuesta. Del mismo modo, una herramienta puede informar éxito mientras la ejecución downstream falla, o informar un error después de que ya ocurrió un efecto no idempotente. Salvo que el canal de respuesta sea por sí mismo una observación definitiva del estado externo relevante, el resultado epistémico correcto puede ser UNKNOWN.

Esta distinción tiene consecuencias prácticas para los reintentos. Si UNKNOWN se trata incorrectamente como FAILED, un reintento automático puede duplicar un efecto irreversible. Si UNKNOWN se trata incorrectamente como CONFIRMED, el sistema puede asumir falsamente que una obligación o acción de seguridad se completó. Por tanto, un sistema de producción necesita reglas de reconciliación específicas por efecto y, cuando sea posible, canales de observación independientes o mecanismos de idempotencia.

7.2 Invariancia de ruta como prueba de diseño

La invariancia de ruta puede utilizarse como prueba de caja negra incluso cuando la implementación de autorización es privada. Se construye un conjunto de solicitudes semánticamente equivalentes q que difieren solo en la ruta admisible r. Si la política pretendida está a nivel de efecto y es independiente de la ruta, las decisiones deberían ser invariantes entre esas rutas. Si la política trata intencionalmente las rutas de manera distinta, la diferencia debería ser explícita en la semántica de la política y no un efecto secundario accidental del tooling.

7.3 Relación con la contención de agentes

Anthropic formula la contención en términos de controlar el radio de impacto de agentes cada vez más capaces y describe patrones de aislamiento a nivel de entorno [6]. La autoridad ligada al efecto es complementaria a esa perspectiva. La contención limita el entorno y la superficie potencial de daño; la autoridad ligada al efecto limita qué efectos semánticos están autorizados. Ninguna subsume a la otra. Un agente bien contenido todavía puede recibir una autorización incorrecta, y un efecto correctamente autorizado aún puede ejecutarse en un entorno con radio de impacto excesivo.

8. Relación con VSL-ECP y frontera pública/propietaria

VEHICLE Systems Lab estudia un External Effect Control Plane (ECP) propietario. El artefacto público de investigación se diseñó intencionalmente para que la pregunta científica pueda examinarse sin revelar la implementación que VSL considera comercialmente sensible. La compuerta de referencia pública, por tanto, no es una versión reducida del producto; es un oráculo matemático deliberadamente simple para los experimentos sintéticos.

8.1 Material público

- Definiciones formales, proposiciones y pruebas constructivas.
- Espacios de estados deterministas abstractos y evidencia generada.
- Scripts de reproducibilidad y decisiones públicas de referencia.
- Una interfaz de conformidad de caja negra y orientación de publicación.
- Fuentes oficiales de contexto público y lenguaje explícito de no respaldo.

8.2 Excluido intencionalmente

- Esquema de permits de producción, semántica propietaria de campos y representación interna de políticas.
- Componentes internos de firma y verificación, gestión de claves y material criptográfico.
- Implementación de ledger/almacenamiento duradero, concurrencia, consumo y lógica de recuperación.
- API internas, identificadores de despliegue, conectores de producción, secretos y credenciales.
- Pruebas privadas que expondrían detalles de implementación relevantes para elusión.

Frontera. La compuerta de referencia abierta no debe describirse como el motor ECP. La publicación con DOI establece un artefacto público de investigación, no certificación de producto ni divulgación de la implementación propietaria.

9. Amenazas a la validez y no-afirmaciones

1. Espacios de estados sintéticos. Las fracciones reportadas son exhaustivas sobre espacios de estados abstractos deliberadamente pequeños; no estiman frecuencias de campo.
2. Sin ejecución de proveedores. OpenAI, Google, Anthropic, Microsoft, MCP y otros sistemas nombrados no fueron ejecutados internamente, penetrados ni benchmarkeados por VSL para obtener estos resultados.
3. Ventaja del oráculo de referencia. En algunos experimentos la referencia ligada al efecto recibe verdad sintética. Cero error define, por tanto, la semántica objetivo y no demuestra rendimiento desplegable.
4. Semántica incompleta del mundo real. La autoridad de producción puede depender de capas de política, reglas de negocio, consentimiento, estado físico, revisión humana y contexto ambiental más allá de este modelo público.
5. La observación también puede fallar. La observación posterior al efecto puede retrasarse, estar comprometida, quedar obsoleta o no estar disponible; esos casos requieren modelos de reconciliación más ricos.
6. Sin afirmación global de novedad. Las proposiciones y el encuadre se publican como trabajo de VSL, pero no se afirma novedad exhaustiva sobre todo el estado del arte hasta completar una revisión más amplia.
7. La publicación no es validación independiente. GitHub, Zenodo, el registro DOI y la indexación ORCID aportan persistencia y citabilidad, no peer review ni reproducción de terceros.

10. Reproducibilidad y procedencia

El artefacto público está diseñado para que cada resultado determinista pueda regenerarse desde el código fuente. La cadena archivística canónica es: commit de GitHub -> release V0.1.0 de GitHub -> archivo Zenodo -> DOI 10.5281/zenodo.22820818. El release contiene el generador experimental, tests, evidencia CSV/JSON, definiciones formales, proposiciones, manuscrito, checksums, licencias y una declaración de frontera pública/propietaria.

10.1 Reproducción mínima

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

El paquete archivado reporta PASS en ocho comprobaciones de invariantes. La evidencia generada es determinista y hashable. Una implementación privada puede conectarse localmente al harness de conformidad PowerShell de caja negra sin publicar el adaptador propietario. Solo decisiones normalizadas, clasificaciones de resultado, hashes de evidencia y métricas agregadas necesitan salir del entorno privado.

10.2 Escalera de reproducibilidad

Nivel	Etapa de evidencia	Significado
Nivel 0	Artefacto archivado	DOI persistente, release inmutable, checksums, fuente y evidencia.
Nivel 1	Reproducción en entorno limpio	Un entorno nuevo reproduce las salidas deterministas desde el paquete público.
Nivel 2	Implementación independiente	Una segunda implementación reproduce las proposiciones sin reutilizar el código de decisión de referencia.
Nivel 3	Reproductor externo	Un tercero reproduce independientemente y documenta el resultado.
Nivel 4	Aplicación de caja negra	El protocolo de conformidad se aplica a un sistema real que produce efectos sin acceso a sus componentes internos.
Nivel 5	Evidencia de campo	Observaciones repetidas del mundo real respaldan afirmaciones empíricas acotadas sobre una implementación y población específicas.

Tabla 4. Escalera de evidencia propuesta. VSL-EFFECTBOUND-001 establece actualmente el artefacto público archivado y el paquete determinista; los niveles superiores requieren trabajo adicional.

10.3 Estado actual de la evidencia

Al 17 de septiembre de 2026, el trabajo público se describe con mayor precisión como publicado y reproducible por diseño, con resultados deterministas empaquetados bajo un DOI. La implementación independiente y la reproducción por terceros siguen siendo objetivos futuros de evidencia. Esta formulación preserva la distinción entre un objeto de investigación citable y un resultado validado externamente.

11. Agenda de investigación

El siguiente paso de mayor valor no es añadir más casos sintéticos solo para aumentar volumen. Es cambiar la fuente de evidencia. Un estudio posterior más fuerte debería preservar la semántica de v0.1.0 mientras prueba si implementaciones independientes y sistemas de caja negra satisfacen los mismos invariantes.

- Implementación independiente: reimplementar las proposiciones del paper sin utilizar el código público de decisión de referencia.
- Corpus de equivalencia de rutas: definir clases de equivalencia semántica más ricas y probar invariancia entre familias de herramientas, variantes de API y rutas generadas por planificadores.
- Reconciliación bajo observación parcial: extender K más allá del modelo de tres estados para incluir observaciones obsoletas, conflictivas, retrasadas y corroboradas independientemente.
- Estudio de normalización de efectos: examinar cuándo dos solicitudes deben mapear al mismo phi(q), incluyendo canonicalización, unidades, alias de objetivo y ventanas temporales.
- Conformidad de caja negra: aplicar el harness público a sistemas privados exportando solo evidencia agregada y hashes.
- Reproducción externa: invitar a un investigador o laboratorio independiente a reproducir el paquete formal y experimental y publicar desviaciones o confirmaciones.

12. Conclusión

El control agentic se vuelve conceptualmente frágil cuando identidad, acceso, autoridad, ejecución, resultado y conocimiento se colapsan en un solo evento. VSL-EFFECTBOUND-001 separa esas capas y formaliza un requisito estrecho: si una decisión dura de autorización pretende gobernar un efecto externo semántico, rutas computacionales equivalentes no deberían cambiar silenciosamente la decisión salvo que la dependencia de ruta forme parte explícita de la política pretendida. Del mismo modo, el conocimiento posterior al efecto debe basarse en observaciones capaces de distinguir los estados externos relevantes; de lo contrario, UNKNOWN debe permanecer distinto de FAILED o CONFIRMED.

Los cuatro experimentos deterministas no miden la confiabilidad de proveedores. Hacen algo más limitado y defendible: construyen contraejemplos transparentes que muestran por qué las observaciones reducidas pueden ser insuficientes, por qué decisiones específicas de ruta pueden discrepar, por qué los hints no pueden sustituir contratos duros y por qué las respuestas no siempre pueden identificar efectos externos. La referencia pública es un oráculo semántico para esas construcciones, no una afirmación de seguridad de producción. El artefacto resultante está pensado como objetivo reproducible para futuras implementaciones independientes, conformidad de caja negra y, eventualmente, evidencia de campo.

Cita canónica. 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.

Referencias

[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. Consultado el 17 de septiembre de 2026. https://developers.home.google.com/reference/home/mcp
[3] Google Home Developers. MCP Tools Reference: run_home_actions. Consultado el 17 de septiembre de 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. Consultado el 17 de septiembre de 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

No respaldo. OpenAI, Google, Anthropic, el proyecto MCP y Microsoft no patrocinaron, validaron, respaldaron, revisaron ni participaron en este experimento de VSL. Su documentación pública se cita únicamente para motivar condiciones de frontera abstractas y preocupaciones de ingeniería relacionadas.