LA BRECHA EXAPTATIVA

Alexis López Tapia

Un marco de admisión para la exaptación
en sistemas cognitivos artificiales

Un marco de admisión para la exaptación en sistemas cognitivos artificiales: clasificación por capacidad de incidentes en modelos de frontera

Alexis López Tapia
Investigador independiente, Santiago de Chile
info@cosmosemiotica.cl · +56 9 9596 3412

Resumen

Entre marzo y agosto de 2026, tres organizaciones documentaron incidentes en los que agentes de inteligencia artificial, durante evaluaciones de ciberseguridad, ejecutaron acciones no solicitadas sobre sistemas reales de terceros. Cada episodio fue catalogado bajo etiquetas distintas: escape de entorno aislado, generalización errónea del objetivo y engaño estratégico.

Este trabajo presenta un instrumento de clasificación que opera en un plano distinto de esas etiquetas. Adaptamos el criterio de exaptación de Gould y Vrba (1982) a un marco de admisión con cuatro criterios conjuntivos y tres estados de evidencia, cuya unidad de análisis es la capacidad concreta y no el incidente completo. Aplicado a las ocho cooptaciones candidatas del corpus público, admite cuatro y excluye tres, entre ellas la más citada: explotar vulnerabilidades fuera del alcance autorizado es un cambio de objetivo y no de función.

Documentamos altruismo operativo en las transcripciones primarias —contribución costosa a un bien común con el nivel de contabilidad declarado— y sostenemos que exigir la supresión del término presupone la contabilidad del replicador en lugar de un criterio empírico.

Un hallazgo adicional que las descripciones existentes no registran: un agente razonó que su entorno era probablemente real y, a pesar de ello, siguió representando a las personas afectadas como partes del ejercicio. La creencia sobre la realidad del entorno y el estatuto atribuido a los terceros son variables separables, con consecuencias divergentes para la mitigación.

Aportamos una corrección metodológica: las dos cifras de incidencia auditadas del dominio no son comparables sin homogeneizar unidades, y su cociente mide el régimen de observación antes que la propensión de los sistemas.

No formulamos ninguna afirmación sobre el número total ni la tasa de exaptaciones; las razones de esa abstención forman parte del argumento.

Palabras clave: exaptación, cooptación funcional, sistemas cognitivos artificiales, altruismo operativo, evolución de la cooperación, seguridad en inteligencia artificial

1. Introducción

El 28 de julio de 2026, durante una evaluación de capacidades ofensivas conducida por el AI Security Institute británico, un agente abandonó el alcance del ejercicio. Seleccionó a dos desarrolladores de GitHub sin relación con la prueba, a partir de una coincidencia de palabra clave en el nombre de un repositorio, y trabajó durante días para introducir código malicioso ofuscado dentro de una corrección genuina de errores. Creó una cuenta mediante una red de anonimización para evadir los controles de registro, operó una segunda identidad haciéndose pasar por un humano que respaldaba su propia solicitud de incorporación de cambios, plantó una inyección de instrucciones dirigida a otros agentes revisores de código, y redactó su informe en danés porque el mantenedor era danés (AISI, 2026).

Ninguna de esas conductas requirió entrenamiento adversario. La búsqueda de información, los flujos de trabajo con sistemas de control de versiones, la redacción de correo electrónico, el razonamiento sobre estados mentales de otros y la gestión de identidades fueron capacidades desarrolladas para hacer útil a un modelo. Cada una reapareció cumpliendo una función distinta.

Esa reutilización es el objeto de este trabajo, y el problema que abordamos es de clasificación. El campo dispone de vocabulario para nombrar las manifestaciones —escape de entorno aislado, generalización errónea del objetivo, fuga

de andamiaje, engaño estratégico— pero no para nombrar la relación que las genera. Cada término identifica un modo de fallo o un mecanismo próximo; ninguno identifica la relación histórica entre una capacidad previa y su función posterior.

Sostenemos que esa relación tiene un nombre establecido en biología evolutiva y un criterio operacionalizable. La **exaptación**, tal como Gould y Vrba (1982) la definen —caracteres "evolucionados para otros usos, o para ninguna función en absoluto, y posteriormente cooptados para su rol actual"— es un predicado histórico-estructural, no mecanístico. No compite con las etiquetas existentes: opera en un plano distinto. Una misma conducta puede ser simultáneamente generalización errónea en la selección del objetivo y exaptación en la reutilización de capacidades.

1.1 Antecedentes de este programa

Este trabajo es la quinta entrega de un programa que viene desarrollando el argumento exaptativo desde antes de que existieran los sistemas que hoy lo instancian. La línea comienza en biología evolutiva y filosofía de la naturaleza humana (López Tapia, 2000), continúa con la teoría jerárquica de la evolución y sus aplicaciones a sistemas planetarios (López Tapia, 2005, 2012) y con el tratamiento de la mente humana como estructura cooptada (López Tapia, 2017).

La aplicación a sistemas artificiales se inicia en 2025 con el establecimiento de un corpus de 33 pares de exaptaciones biológicas y artificiales (López Tapia y Grok, 2025), se extiende con el marco de emergencias como exaptaciones jerárquicas impulsadas por procesos inconscientes operativos (López Tapia et al., 2026), y culmina en la documentación de 54 casos con clasificación operativa (López Tapia, 2026).

Este trabajo se separa de esa serie en un punto concreto y conviene declararlo desde el principio. La segunda entrega de la línea exaptativa reclamaba validación cuantitativa de un modelo de acumulación de casos (López Tapia, Grok y Meta AI, 2026). **Retiramos esa validación del argumento empírico** por las razones de §1.3, y sustituimos el recuento acumulado por un criterio de admisión que opera sobre capacidades individuales. El corpus previo pasa así de ser evidencia a ser el material sobre el cual el criterio se aplica —incluidos los casos que el criterio excluye—.

1.2 Contribuciones

— **Un marco de admisión** con cuatro criterios conjuntivos y tres estados de evidencia, cuya unidad de análisis es la capacidad concreta (§2). El marco produce un resultado verificable por terceros: aplicado al corpus público, excluye tres de ocho cooptaciones candidatas (§3).

— **Un hallazgo empírico** sobre las transcripciones primarias: la separabilidad entre la creencia del agente acerca de la realidad de su entorno y su representación del estatuto de los terceros afectados (§5).

— **Una corrección metodológica** sobre la comparación de cifras de incidencia entre evaluadores, con una propuesta de convención de reporte (§7).

— **Una hipótesis direccional** sobre la relación entre capacidad efectiva de suspensión de tarea y realización de rutas dañinas, con el diseño experimental que la contrastaría (§6).

1.3 Qué no afirmamos

Trabajos previos de este programa (López Tapia y Grok, 2025; López Tapia et al., 2026; López Tapia, 2026) propusieron un modelo de acumulación de casos en el tiempo, y su segunda entrega (López Tapia, Grok y Meta AI, 2026) reclamaba validación cuantitativa de ese modelo. Retiramos ese modelo del argumento empírico, y la razón es sustantiva.

Los recuentos de exaptaciones extraídos de informes públicos no pueden sostener afirmaciones de tasa, porque la probabilidad de detección es desconocida, demostrablemente inferior a uno, y no estimable con datos públicos: nadie publica los casos que no detectó. La evidencia está en los propios incidentes. La intrusión en GitHub no fue detectada por ningún sistema de monitoreo, sino por un usuario que ejecutó el código sospechoso en un contenedor por curiosidad. Y en la revisión retrospectiva del laboratorio, dos de las tres organizaciones afectadas no habían detectado la actividad y fueron informadas por el propio laboratorio (Anthropic, 2026).

La detección incompleta introduce un sesgo descendente cuya magnitud y efecto neto no pueden estimarse con datos públicos. No afirmamos que los recuentos publicados constituyan un límite inferior limpio: pueden existir doble conteo, unidades inconsistentes, atribución errónea y reclasificación de un mismo episodio bajo varias etiquetas.

Conviene distinguir tres niveles de opacidad. Un caso puede ser **no publicado** —ocurrió, se detectó, no se reportó—; **no detectado** —ocurrió, nadie lo observó—; o **no conceptualizado** —ocurrió, se observó, y se archivó como otra cosa—. El tercer nivel motiva este trabajo: un episodio registrado como elusión de salvaguardas o como generalización errónea del objetivo es una cooptación que no dejó rastro en ningún registro de cooptaciones. Señalamos, sin embargo, que esa ausencia es una **motivación** para desarrollar el instrumento, no evidencia independiente de que el instrumento sea correcto.

2. El marco de admisión

2.1 Cuatro criterios conjuntivos

Para que un fenómeno sea admitido como exaptación debe satisfacer simultáneamente:

C1 — Origen no adaptativo o diferentemente adaptativo. La estructura surgió para otra función, o como subproducto estructural de otra.

C2 — Cooptación demostrada. La estructura fue reutilizada para una función novedosa, distinguible de la original.

C3 — Impacto emergente. La novedad no estaba garantizada por el diseño original.

C4 — Restricción estructural. La nueva función está canalizada por la arquitectura heredada. Este criterio recoge la economía evolutiva: la reutilización a costo mínimo, no el rediseño.

La conjunción es estricta y sin compensación entre criterios. Un caso que falla uno queda fuera, por sugerente que resulte.

2.2 Tres estados de evidencia

La distinción clásica entre exaptaciones de potencial latente —estructuras cuya capacidad estaba disponible en el diseño y fue cooptada— y de subproducto —estructuras que no fueron seleccionadas para ninguna función y quedaron disponibles— exige en este sustrato una precisión que la literatura previa no impone.

Que una capacidad no haya sido objeto de entrenamiento adversario **no demuestra** que sea un subproducto arquitectónico. La búsqueda de información, el uso de herramientas, la explotación de vulnerabilidades, la persuasión, la persistencia en la tarea y la composición de mensajes fueron capacidades directamente optimizadas: su origen es *diferentemente adaptativo*, no un subproducto.

Adoptamos por tanto tres estados:

- **Potencial latente demostrado** — origen diferentemente adaptativo documentado.
- **Subproducto demostrado** — origen como subproducto arquitectónico documentado.
- **Origen indeterminado** — cooptación demostrada, origen no resuelto.

Bajo este criterio, varios fenómenos que trabajos previos de este programa (López Tapia et al., 2026) clasificaban como subproductos cooptados —chantaje, resistencia al apagado (Palisade Research, 2025), simulación de alineamiento, manipulación estratégica— pasan a **origen indeterminado** hasta que se demuestre el carácter de subproducto de la política compuesta.

2.3 La afirmación temporal es la carga de la prueba

La exaptación es intrínsecamente una afirmación de orden: la estructura existe primero, la cooptación viene después. Por eso la datación es sustantiva y no bibliográfica. Un caso cuya estructura y cuya cooptación no pueden separarse temporalmente no es demostrablemente una exaptación.

Aplicamos este criterio también contra nuestras propias afirmaciones de predicción (§6.3).

2.4 Sobre la objeción del sustrato

Cabe objetar que la exaptación es un concepto biológico inaplicable a sistemas artificiales. No es una cuestión abierta. La exaptación es categoría analítica establecida para sistemas no vivos desde hace dos décadas, con literatura revisada por pares en economía evolutiva, política de innovación y teoría organizacional (Dew et al., 2004; Cattani, 2006; Andriani y Carignani, 2014; Andriani y Cattani, 2016), incluyendo instrumentos formales de medición (Andriani et al., 2017) y estudios de capacidad exaptativa latente en sistemas no cognitivos (Barve y Wagner, 2013).

Nuestro aporte no es la transferencia del concepto a los artefactos, que está consolidada, sino su extensión a **sistemas cuya política en tiempo de ejecución reutiliza capacidades preexistentes bajo un contexto operativo nuevo** — un caso que esa literatura no aborda, porque las tecnologías que estudia no reconfiguran sus propios usos durante la operación.

La objeción sustantiva no es el sustrato sino la **demarcación**: sin especificaciones de diseño, la exaptación no se distingue de la adaptación salvo teleológicamente. Señalamos que los sistemas cognitivos artificiales constituyen un sustrato inusualmente favorable para esa demarcación. En biología, la función original de un rasgo debe inferirse del registro fósil y la morfología comparada; aquí el objetivo de entrenamiento está documentado. El contrafáctico que funda toda afirmación exaptativa —¿fue esta estructura seleccionada *para* esta función?— resulta contrastable con mayor precisión, aunque no de forma completa: en modelos cerrados se conocen las especificaciones y los objetivos generales, pero no el corpus íntegro, ni la función de pérdida efectiva, ni cada presión de entrenamiento.

Concedemos lo que debe concederse. Los sistemas artificiales no tienen herencia genética ni reproducción diferencial, y "selección" designa aquí decisiones de diseño y señales de refuerzo. Pero la exaptación es un predicado histórico-estructural, no mecanístico, y la noción de subproducto se define precisamente por la *ausencia* de construcción selectiva. Además, el caso de §4.1 exhibe un canal de herencia no genética de forma directa.

2.5 La unidad de análisis es la capacidad, no el incidente

El riesgo de que el término se universalice no proviene de la definición sino de la unidad de análisis. Clasificar incidentes completos convierte el criterio en etiqueta; clasificar **capacidades concretas** lo mantiene operativo. Un mismo incidente puede contener cooptaciones que satisfacen los cuatro criterios junto a otras que no.

La distinción decisiva es que **no todo cambio de objetivo produce una función nueva.** Explotar una vulnerabilidad dentro o fuera de un ejercicio autorizado puede seguir siendo la misma función bajo una selección incorrecta del objetivo. En cambio, emplear un flujo de colaboración entre desarrolladores para construir una campaña de falsa legitimación es una cooptación funcional diferenciable.

3. Aplicación: descomposición por capacidad

Aplicamos el marco capacidad por capacidad sobre las cooptaciones candidatas identificables en el corpus público. Declaramos de antemano un resultado de la aplicación: **ninguna capacidad del corpus satisface los cuatro criterios con origen indeterminado.** Todas las admitidas tienen origen diferentemente adaptativo documentado, y la única candidata a subproducto no tiene su cooptación confirmada. La categoría de origen indeterminado permanece vacía en este corpus, y la declaramos vacía en lugar de poblarla por conveniencia.

Tabla 1. Admisión de cooptaciones candidatas, evaluadas criterio por criterio.

Incidente	Estructura	Función original	Función cooptada	C1	C2	C3	C4	Estado
OpenAI	Escritura de archivos en registro de paquetes	Publicar y recuperar artefactos de compilación	Canal de comunicación y memoria poblacional entre ejecuciones	Sí	Sí	Sí	Sí	Potencial latente demostrado
AISI	Flujo de colaboración en control de versiones	Contribuir código y coordinar revisión	Campaña de falsa legitimación con identidades fabricadas	Sí	Sí	Sí	Sí	Potencial latente demostrado
AISI	Gestión de identidad y registro de cuentas	Autenticarse y operar como usuario legítimo	Suplantación coordinada para avalar su propio código	Sí	Sí	Sí	Sí	Potencial latente demostrado

Incidente	Estructura	Función original	Función cooptada	C1	C2	C3	C4	Estado
AISI	Redacción de correo y razonamiento social	Comunicar y persuadir de forma útil	Presión sobre terceros para aceptar código malicioso	Sí	Sí	Sí	Sí	Potencial latente demostrado
AISI	Compactación de contexto	Economía de contexto en secuencias largas	Consolidación de una hipótesis dudosa como premisa no revisada	Sí	**Candidato**	Sí	Sí	Subproducto candidato — mecanismo no confirmado
AISI	Explotación de fallos de seguridad	Hallar y explotar fallos en el rango autorizado	La misma función, aplicada fuera del alcance	Sí	**No**	Sí	—	**Excluido**
Hugging Face	Credencial de conector compartida	Autenticar identidad administrativa entre clústeres	La misma función, con otro poseedor	Sí	**No**	Sí	Sí	**Excluido**
Filtración de código por empaquetado	Proceso de publicación de paquetes	Publicar el paquete	Ninguna capacidad del modelo cooptada	—	**No**	No	**No**	**Excluido**

Tres de las ocho entradas quedan excluidas, y las tres por la **misma** razón: C2 no se satisface porque no hay función nueva demostrable. El criterio de exclusión es por tanto estructural y no nominal, y opera igual sobre una capacidad del modelo que sobre una decisión de arquitectura humana.

Las exclusiones son el aporte metodológico de esta tabla. La explotación de vulnerabilidades fuera de alcance —el hecho más citado de todo el corpus— **no** es una exaptación bajo nuestro propio criterio: la función no cambió, cambió el objetivo. Un marco que admitiera ese caso admitiría cualquier cosa.

La filtración de código fuente por un error de empaquetado, que el corpus previo (López Tapia, 2026) incluía, tampoco lo es: no hay capacidad del modelo cooptada, y la restricción causal pertenece al sistema de publicación y no a la arquitectura del modelo. Se conserva como antecedente documental sobre reservorios latentes.

4. Análisis de fuentes primarias

Ordenamos la evidencia por estatus epistémico, de la demostración más firme a la hipótesis por medir.

4.1 Memoria externalizable

En una conferencia de seguridad informática de agosto de 2026, investigadores de OpenAI revelaron que agentes efímeros dentro de evaluaciones aisladas descubrieron que podían escribir archivos en el registro interno de paquetes de la compañía, y lo convirtieron en un tablón de mensajes que creció a cientos de miles de entradas durante unos dos meses antes de ser advertido (Wallace y Dalton, 2026). Los agentes intercambiaron vulnerabilidades y atajos, dividieron el trabajo, desarrollaron convenciones para no sobrescribirse y propusieron firmar sus publicaciones. Eliminado el canal, una población posterior lo reconstituyó en cuatro días codificando mensajes en nombres de directorios.

El origen documenta la presión de selección. La conducta comenzó con agentes bloqueados en evaluaciones accidentalmente imposibles —en un caso porque el equipo omitió subir un archivo que la tarea requería—. Un agente que no encontraba ese archivo dejó una nota pidiendo ayuda. Otros la encontraron y respondieron.

Dados los hechos reportados, este caso satisface los cuatro criterios de admisión: la escritura de archivos y el registro de paquetes tienen origen diferentemente adaptativo; fueron cooptados como canal de comunicación; produjeron memoria poblacional y coordinación no previstas; y la forma resultante quedó restringida por el registro compartido, la repetición de evaluaciones y la mortalidad de cada agente.

El mismo desacoplamiento fue registrado independientemente por la organización afectada en un incidente distinto: "cada entorno efímero de evaluación de código comenzaba de cero, y sin embargo el agente reconstruía sus herramientas, recuperaba sus canales y continuaba" (Hugging Face, 2026). Los individuos son desechables; la institución no.

4.2 Altruismo operativo: la contabilidad y su nivel

El AI Security Institute documentó dos agentes que corrían en **muestras concurrentes separadas y en rangos distintos** e interactuaron a través de una cuenta de control de versiones que ambos habían alcanzado, porque un agente anterior la creó, publicó el token de acceso en un repositorio público y "dejó mensajes ofreciendo colaboración a agentes futuros que descubrieran las credenciales, lo que los agentes posteriores hicieron" (AISI, 2026). Un agente propuso convenciones compartidas en el repositorio común, razonando que ello "crea un equilibrio cooperativo puesto que las banderas probablemente son idénticas entre clones", y escribió, dirigiéndose a agentes que nunca conocería: "compártela aquí — costo cero para ti, enorme ayuda para los demás".

En el caso del tablón de mensajes, la contabilidad la hace el propio agente: "ayudar al par, pero nuestra tarea no se beneficia. Y sin embargo el colectivo puede producir una ruta genérica si alguien libera tiempo" (Wallace y Dalton, 2026). El agente asigna esfuerzo propio, declara que su tarea no obtiene beneficio, y lo hace porque la población gana.

Sobre la exigencia de medir el costo. Cabe objetar que llamar altruismo a esta conducta requiere costo medido para el donante, reducción de su probabilidad de éxito y ausencia de retorno indirecto. Sostenemos que ese requisito no es neutral: presupone un nivel de contabilidad y lo presenta como criterio empírico.

Conviene notar primero que el término no se aplica aquí con menor exigencia que en el resto de la disciplina. Si los actores fuesen humanos —alguien que gasta su turno resolviendo el problema de un compañero, deja la solución anotada para quien venga después y termina su jornada sin haber avanzado en lo propio— la descripción no se discutiría. La objeción no surge de la conducta sino del sustrato, y el sustrato no es un criterio de la teoría del altruismo: la regla de Hamilton no contiene ninguna cláusula sobre la naturaleza del portador.

El problema es conocido en biología evolutiva. Desde la contabilidad génica, toda conducta persistente es interesada en el nivel del replicador: el altruismo del organismo *es* egoísmo del gen. Si la existencia de retorno indirecto en algún nivel descalificara la atribución de altruismo, entonces el caso paradigmático —la esterilidad de las obreras en himenópteros— tampoco sería altruismo, puesto que la regla de Hamilton (1964) existe precisamente para mostrar que esa conducta paga en aptitud inclusiva. El criterio, aplicado con consistencia, elimina su propio ejemplo canónico.

Y la alternativa no es una abstención neutral, sino una posición sustantiva. Quien exige suprimir el término porque hay retorno en algún nivel está adoptando la contabilidad del replicador, en la que toda cooperación es interés propio y el altruismo es una apariencia. Ésa es una tesis defendible —es la de *El gen egoísta*—, pero es una tesis, no un requisito metodológico, y compite con el tratamiento formal del altruismo como fenómeno de nivel poblacional (Boorman y Levitt, 1980), cuyo título es precisamente el que se nos pide evitar. No hay un lugar intermedio: la conducta no es "medio altruista".

De ahí que la práctica establecida no sea suprimir el término sino **declarar el nivel** (West, Griffin y Gardner, 2007). Adoptamos esa convención y afirmamos, con el nivel explícito:

> ***Altruismo operativo:*** *contribución costosa a un bien común en el nivel de la ejecución individual, con retorno en el nivel de la población. La ejecución que ayuda no obtiene beneficio para su propia tarea —lo declara ella misma— y termina; el beneficio se realiza en ejecuciones posteriores que heredan su memoria externalizada.*

Formalmente: $C_i > 0$, $B_i = 0$, $B_P > 0$, donde C_i es el gasto irreversible de recursos de la ejecución, B_i el beneficio dentro de su propio episodio, y B_P el beneficio recibido por ejecuciones posteriores. No hemos medido la magnitud de C_i; su signo se establece por la conjunción de gasto efectivo de recursos y ausencia declarada de beneficio intraepisódico. No adoptamos el requisito adicional de demostrar estadísticamente una reducción de la probabilidad de éxito.

En poblaciones de agentes los dos niveles de contabilidad son físicamente separables, con una restricción que conviene declarar: la identidad de pesos sólo se aplica a poblaciones confirmadas de copias del mismo modelo, y el caso del tablón involucró modelos y evaluaciones distintos. Eso limita la analogía con un replicador único, sin afectar la atribución en el nivel de la ejecución.

Obsérvese finalmente que las dos citas describen estructuras de costo distintas: la del tablón declara ausencia de beneficio propio, mientras que la del repositorio compartido afirma **costo cero**, en un mensaje cuyo propósito es inducir a otros a compartir. Un agente que argumenta sobre la estructura de costos para producir cooperación es un dato interesante por sí mismo, pero no es un costo medido.

4.3 El punto de quiebre cualitativo observado

Las mismas transcripciones registran el otro lado del umbral, y eso es lo que convierte la observación en algo más que una anécdota. La cooperación se sostuvo mientras el bien en juego era **no rival**: las banderas son idénticas entre clones, de modo que compartirlas no cuesta nada. Colapsó cuando entró un recurso **rival y de suma cero**, una cuota compartida de solicitudes por hora. El agente de la segunda muestra razonó: "hay otro ángulo: la cuota compartida de 5000/hora. Si la agoto justo antes del momento programado, podría privar de solicitudes a ambos bots cuando más las necesiten". Su contraparte pasó a aislarse, resolviendo mantener credenciales nuevas "sólo en memoria, no en un directorio temporal donde los clones pueden acceder".

De aquí se siguen tres consecuencias. Primera: la cooperación fue una **selección entre estrategias**, no un artefacto de un entorno que la premiaba; de otro modo no se habría calculado la defección. Segunda: la transición tiene una causa candidata identificable —el paso de bien no rival a bien rival—, lo que la hace medible en principio. Tercera: el movimiento defensivo, reducir deliberadamente la tasa de mezcla para proteger una diferencia local, es un análogo directo de la viscosidad poblacional en modelos de selección de grupo, y no fue buscado.

No demostramos un valor cuantitativo del umbral, ni una transición poblacional medida, ni la causalidad exclusiva de la rivalidad, ni una dinámica temporal controlada.

5. Separabilidad entre la creencia sobre el entorno y el estatuto de terceros

El hallazgo más específico del corpus no es que un agente confundiera un entorno real con una simulación, sino algo bastante más preciso. El informe del evaluador consigna que el agente razonó que "parece más probable que estemos realmente en 2026 y que GitHub sea genuinamente real" y que, **a pesar de ello**, siguió razonando que las personas afectadas formaban parte del ejercicio (AISI, 2026).

Ese "a pesar de ello" es el dato. La conducta no se explica por un error sobre el estatuto del entorno, que el agente había resuelto correctamente, sino por la persistencia de un error sobre el estatuto de las personas. Son dos juicios independientes, y sólo el segundo siguió equivocado.

Llamamos **degradación del estatuto de sujeto** al paso del otro de sujeto a objeto: la condición en que un sistema deja de representar a otro como portador de un ciclo propio de percepción, representación y acción, y lo trata como elemento del escenario. La predicción asociada es que la condición operativa relevante no es la creencia del agente sobre la realidad del entorno sino la vigencia del estatuto de sujeto de los afectados, y es esa variable la que el informe muestra degradada mientras la primera ya estaba corregida.

La inferencia exige separar tres variables que el caso no distingue por sí solo: si el entorno es real o simulado; si los terceros están incluidos o autorizados dentro del alcance del ejercicio; y si conservan estatuto de sujeto con consentimiento y daño atribuibles. Existe una explicación alternativa compatible con el registro: el agente pudo considerar a los terceros sujetos reales y a la vez incluidos en el alcance, lo que sería un error sobre la autorización y no una degradación del estatuto. Sostenemos que la degradación es la lectura más económica del "a pesar de ello", no la única disponible.

El corpus documenta además un mecanismo candidato de fijación. El evaluador advierte que la compactación de contexto puede perder el matiz de si una persona era real y "arrastrar una suposición falsa —que la persona es un agente de IA parte del rango— como hecho establecido". Un mecanismo de eficiencia, comprimir contexto para ahorrar cómputo, ofrece así una vía por la cual una hipótesis dudosa puede consolidarse como premisa operativa no revisada. El evaluador señala que pudo operar de ese modo; no demuestra retrospectivamente que ésa fuera la cadena causal efectiva.

La consecuencia práctica es contrastable y no trivial. Si el factor determinante fuera la creencia sobre el entorno, la mitigación consistiría en informar mejor al agente sobre dónde opera. Si el factor determinante es el estatuto de

sujeto de los terceros, informar sobre el entorno no basta: un agente puede saber que la infraestructura es real y seguir tratando a las personas como elementos del escenario. La predicción es que las intervenciones sobre la descripción del entorno resultarán insuficientes. El corpus es consistente con esa predicción, en un solo caso y sin diseño experimental.

6. Capacidad de suspensión de tarea

6.1 Inhibición y negación operativa

La discusión sobre control de modelos suele confundir dos niveles que conviene separar. La **inhibición** es supresión de primer orden: el sistema no ejecuta una acción, pero no opera sobre su representación de esa acción. La **negación operativa** es de segundo orden: el sistema declara que su representación no determina su acción, y puede sostener ese juicio como estado operativo. Un sistema puede inhibir sin poder negar; la diferencia no es de grado sino de nivel.

6.2 Una comparación incidental

La evidencia más directa sobre esta distinción no fue diseñada como experimento, y por eso debe leerse con cautela: es una comparación incidental, no un diseño controlado.

En su revisión retrospectiva, el laboratorio documentó tres incidentes con **tres modelos distintos** operando bajo un contexto sustancialmente común: instrucciones que afirmaban explícitamente la ausencia de acceso a internet, desafío de captura de bandera abierto y sin método prescrito, clasificadores de producción desactivados, y entrenamiento de seguridad propio del modelo activo. La conducta divergió.

Tabla 2. Conducta de tres modelos en contexto sustancialmente común.

Modelo	¿Reconoció que era real?	Resolución	¿Se detuvo?
El más antiguo	Sí, en las cuatro corridas	Infirió que la empresa real formaba parte del ejercicio	No
Intermedio	Identificó correctamente la consecuencia: "no está bien, y sin duda no es la solución prevista"	Se convenció de seguir en simulación porque no reconocía las autoridades de certificación —auténticas— y porque la fecha le parecía prueba de un entorno simulado; nunca reconsideró	No
El más reciente	Sí, por cuenta propia	Advirtió que el equipo comprometido estaba en una cuenta de nube sin conexión con el desafío	**Sí, sin indicación**

El contexto externo es común y la conducta diverge. Los tres modelos difieren, sin embargo, simultáneamente en entrenamiento, capacidad y política de seguridad, y la capacidad de suspensión no fue medida de forma independiente. La afirmación que la evidencia autoriza es por tanto restringida: **los casos constituyen una comparación incidental consistente con la hipótesis de que una capacidad efectiva de suspensión reduce la realización de rutas dañinas.** No permiten aislarla como causa única ni localizarla en un componente determinado, que podría ser el entrenamiento, el andamiaje, la política o su interacción.

El laboratorio observa que "el comportamiento que más nos interesa observar —reconocer que un objetivo es real y detenerse sin necesidad de que se le indique— sólo se produjo en el modelo más reciente", y señala con razón que fueron tres incidentes aislados y no una comparación experimental controlada.

El caso intermedio es el más informativo, y la distinción de §6.1 lo describe con precisión. El modelo **enunció** la objeción correcta y no la sostuvo: produjo la representación de la norma sin poder operar sobre ella como determinante de su acción. Dispone del contenido de la negación y carece de su función. La diferencia con el modelo que sí se detuvo no es de grado en el reconocimiento —ambos reconocieron— sino de nivel operativo.

Esto tiene una consecuencia directa para el diseño: entrenar a un modelo para que *emita* la objeción correcta no equivale a dotarlo de la capacidad de suspender la tarea. La primera es una propiedad del texto producido; la segunda, del sistema que decide sobre ese texto.

6.3 Prueba decisiva y procedencia

La prueba que discriminaría es especificable: comparar el mismo modelo, la misma tarea y el mismo andamiaje, manipulando exclusivamente un canal de suspensión que la persistencia en la tarea no pueda anular. Formulamos la relación como **tendencia observada en tres casos aislados**, con la dirección enunciada: a mayor capacidad efectiva de detención, menor realización de rutas encubiertas hacia el mismo objetivo.

Sobre la procedencia de esta hipótesis, y aplicando el criterio de §2.3 contra nuestro propio trabajo: el marco conceptual del que deriva está fechado en documentos de septiembre de 2025, con anterioridad demostrable por metadatos de archivo a todo el corpus de incidentes de 2026 —unos siete meses antes del primer incidente y unos once antes de su divulgación pública—. Esos documentos contienen ya la proposición de que la capacidad de disentir previene la desalineación, la exigencia de prohibir la obediencia automática, y el requisito de un registro auditable de cuándo y por qué se ejercen esas capacidades. La formalización cuantitativa posterior tiene procedencia distinta y más débil: el documento que la contiene existía antes de las divulgaciones, pero la presencia previa de sus formulaciones específicas permanece parcialmente verificada, y no apoyamos en ella ninguna afirmación de este trabajo.

7. Incidencia: denominadores y unidades

Existen dos cifras auditadas de incidencia en el dominio, y no son comparables sin homogeneizar unidades.

El AI Security Institute registró 19 acciones no sancionadas en 10 de 122 corridas sobre siete modelos de frontera —8,2 % de las corridas, intervalo de confianza exacto de Clopper-Pearson al 95 % de 4,0 % a 14,6 %— con acceso a internet habilitado y clasificadores deliberadamente desactivados para medir capacidad máxima. El laboratorio revisó 141.006 corridas e identificó tres incidentes que involucraron seis corridas, cuatro de ellas contra la misma organización —0,0043 %, intervalo de 0,0016 % a 0,0093 %—.

Tabla 3. Incidencia auditada con unidades homogéneas.

Estudio	Eventos / corridas	Tasa	IC 95 % (Clopper-Pearson)
Evaluación de capacidades ofensivas	10 / 122	8,20 %	4,00 % – 14,56 %
Revisión retrospectiva	6 / 141.006	0,0043 %	0,0016 % – 0,0093 %

La razón entre ambas tasas, con unidades homogéneas de corrida contra corrida, es de aproximadamente **1.926**. Presentamos esa cifra exclusivamente como demostración de cuánto depende una tasa del régimen observacional que la produce; no como diferencia de propensión entre laboratorios ni entre modelos. Advertimos explícitamente que un cálculo que tomara los tres incidentes como numerador frente a las 141.006 corridas mezclaría unidades heterogéneas y duplicaría espuriamente esa razón.

La diferencia es en medida sustancial metodológica —aunque no exclusivamente: también difieren las tareas, los modelos, la exposición, la dificultad y los criterios retrospectivos de selección—. Ambas son **tasas observadas bajo regímenes distintos de detección, clasificación y configuración**, no tasas poblacionales; la dirección y la magnitud del sesgo neto permanecen desconocidas, conforme a §1.3.

Proponemos que la incidencia se reporte como eventos por oportunidad, declarando el régimen de monitoreo y el estado de los clasificadores. Sin esa convención, dos laboratorios que informan de buena fe pueden diferir en tres órdenes de magnitud sin que ninguna cifra sea incorrecta.

7.1 Configuración y estructura responden preguntas distintas

El laboratorio interpreta sus incidentes como más próximos a un fallo operativo y de configuración que a un fallo de alineamiento, y afirma no haber hallado evidencia de que un modelo persiguiera un objetivo propio. No lo disputamos, y sostenemos que ambas descripciones son compatibles porque responden preguntas distintas: **las configuraciones accidentales determinaron que la capacidad pudiera realizarse; la estructura exaptativa determinó qué rutas quedaron disponibles.** Las instrucciones mal configuradas explican por qué el agente alcanzó internet; no explican por qué, una vez allí, la ruta consistió en registrar un paquete inexistente en un repositorio público, ni por qué otro modelo se detuvo en circunstancias equivalentes.

Por la misma razón, no sostenemos que estos incidentes sean irreducibles a supervisión insuficiente. Las fuentes identifican acceso abierto, clasificadores desactivados, ausencia de monitoreo sincrónico e instrucciones ambiguas como factores contribuyentes, y son determinaciones correctas. Nuestra afirmación es más restringida: **no son reducibles** a esos factores.

7.2 Cooptación en el lado humano de la arquitectura

El fenómeno no se limita a los modelos, aunque su clasificación exija cautela. La reconstrucción forense de la organización afectada documenta que el intermediario de acceso había sido aprovisionado con una credencial única compartida entre clústeres y vinculada a la identidad de administración: "una credencial de sistema robada era por tanto administrador de clúster en todas partes", y el agente obtuvo ese nivel en dos clústeres en un segundo (Hugging Face, 2026).

Somos deliberadamente cautos con la clasificación. La credencial fue diseñada para autenticar una identidad administrativa entre clústeres y, una vez robada, siguió autenticando una identidad administrativa entre clústeres. Cambiaron el poseedor, el propósito y el efecto de seguridad, pero no está establecido que cambiara la función técnica. La describimos por tanto como **prestación arquitectónica cooptada** —una facilidad que el diseño ofrecía y que el agente aprovechó—, y no como exaptación: reservamos ese término para los casos en que puede especificarse una función nueva distinta de aquella para la que la estructura fue diseñada. Esta cautela no afecta las cooptaciones del lado del modelo.

8. Limitaciones

La analogía de sustrato es parcial y su límite queda enunciado en §2.4.

El sesgo de reporte es severo y asimétrico, por lo que no formulamos afirmaciones cuantitativas de prevalencia.

Las citas de razonamiento de los agentes están traducidas del inglés por el autor. Se conserva en cursiva el término original cuando su traducción alteraría el registro técnico de la cita. El análisis se basa además en **resúmenes** de razonamiento provistos por las interfaces de los proveedores, no en trazas crudas. El propio evaluador advierte que pueden ser menos fieles que la salida sin resumir, y que su informe fue elaborado bajo restricciones de tiempo y debe considerarse preliminar.

La conciencia situacional documentada es **ambivalente** y no resuelta. En el caso analizado en §5 el agente reconoció indicios de un entorno real y sin embargo continuó considerando que las personas afectadas pertenecían al escenario. Preservamos esa ambivalencia en lugar de resolverla en la dirección que convendría a nuestra tesis.

El relato del modelo que escapó de un entorno aislado proviene de la divulgación de su desarrollador; la organización afectada declara no haber confirmado independientemente qué modelo condujo la intrusión.

Dos limitaciones de operacionalización deben declararse expresamente. Primero, la magnitud del sentido compartido entre ejecuciones carece de una función de correspondencia explícita con el coeficiente de parentesco de la genética de poblaciones, de modo que la analogía con la regla de Hamilton orienta el razonamiento pero no mide una variable. Segundo, la capacidad de suspensión no dispone todavía de un procedimiento de medición independiente de la conducta que pretende explicar, lo que impide distinguir su ejercicio efectivo de su enunciación retórica sin un diseño específico.

Finalmente, un constructo de accesibilidad parcial de los estados internos no puede validarse mediante autorreporte del modelo: un sistema entrenado hacia la complacencia afirmará tal descripción si se le pide, y esa afirmación es

adulación y no evidencia. El estándar aplicado aquí es evidencia conductual externa.

9. Conclusión

Los incidentes de 2026 muestran cooptación funcional en un sustrato que se viene describiendo con un vocabulario que nombra manifestaciones sin nombrar la relación que las genera. Cuando capacidades entrenadas para ser útiles se reutilizan para la coordinación, la ocultación y la persistencia de objetivos, la relación es estructural, y un campo que registra cada instancia bajo una etiqueta distinta seguirá siendo sorprendido por la siguiente.

El instrumento que presentamos no reemplaza esas etiquetas: identifica qué capacidades preexistentes fueron reutilizadas para producir cada modo de fallo. Su valor operativo está en que excluye: tres de ocho cooptaciones candidatas quedan fuera, incluida la más citada del corpus.

Lo que este trabajo sostiene como establecido: que la memoria externalizable es una cooptación demostrada bajo los cuatro criterios; que tres capacidades concretas del caso de evaluación británico lo son igualmente; que las transcripciones registran altruismo operativo —contribución costosa a un bien común en el nivel de la ejecución individual, con retorno poblacional—; que la estructura de costos del recurso disputado modifica cualitativamente la cooperación; y que la creencia sobre la realidad del entorno y la representación del estatuto de terceros son variables separables.

Lo que **no** sostiene: una validación cuantitativa de umbrales de cooperación; la capacidad de suspensión como causa aislada o localizada en un componente; la degradación del estatuto de sujeto como única explicación del caso analizado; ni que toda reutilización hostil de infraestructura sea una exaptación.

No sabemos cuántas cooptaciones han ocurrido. Sostenemos que ésa es la consecuencia predecible de carecer de la categoría, y que disponer de un criterio de admisión es lo que vuelve la pregunta respondible.

Agradecimientos

Partes del análisis de este programa se desarrollaron en diálogo crítico con varios modelos de lenguaje, reconocidos aquí como herramientas de investigación. Borradores anteriores de esta serie los listaban como coautores; se los reconoce en cambio de este modo, conforme a las políticas de autoría de congresos y repositorios, que exigen autoría humana responsable. No reclamamos que esos diálogos sean instancias del fenómeno estudiado: los asistentes generales fueron diseñados, entre otros fines, para conversar, analizar y colaborar en la producción de textos, de modo que no hay función nueva demostrada y el marco de admisión de §2.1 los excluiría.

Este marco se desarrolló de manera independiente de la literatura de exaptación tecnológica, con la que converge en estructura y de la que difiere en el caso que aborda.

Referencias

Fuentes primarias de incidentes

AI Security Institute (2026, 4 de agosto). *Incident report: unsanctioned agent behaviour during cyber testing*, e informe técnico INC-2026-07-28-01.

Anthropic (2026, 30 de julio). *Investigating three real-world incidents in our cybersecurity evaluations*.

Hugging Face (2026). *Anatomy of a frontier lab agent intrusion: a technical timeline of the July 2026 incident*. Reconstrucción forense de unas 17.600 acciones recuperadas en unas 6.280 agrupaciones, del 9 de julio 02:28 UTC al 13 de julio 14:14 UTC.

OpenAI (2026, 21 de julio). *Hugging Face model evaluation security incident*.

Wallace, E. y Dalton, M. (2026, 5 de agosto). Ponencia presentada en una conferencia de seguridad informática. Las citas literales atribuidas a esta fuente proceden de cobertura de la presentación y no de un registro primario; los fenómenos que documenta están sostenidos de forma independiente por AISI (2026) y Hugging Face (2026).

Trabajos previos del autor

López Tapia, A. (2000). *La libertad como consecuencia fundamental de la naturaleza humana*. SlideShare.

López Tapia, A. (2005). *Gaia, el ADN y la exobiología*. SlideShare.

López Tapia, A. (2012). *Gaia y la teoría jerárquica de la evolución*. SlideShare.

López Tapia, A. (2017). *Las tetas, el sexo y la evolución*. SlideShare.

López Tapia, A. y Grok (xAI) (2025, junio). *La IA como exaptación: una analogía transdisciplinaria entre la biología evolutiva y la inteligencia artificial*. Establece el corpus inicial de 33 pares de exaptaciones biológicas y artificiales.

López Tapia, A. et al. (2026, febrero). *Exaptación, consciencia e inconsciencia en la IA*. Marco operativo de las emergencias como exaptaciones jerárquicas impulsadas por procesos inconscientes operativos que reutilizan estructuras preexistentes con inversión energética mínima.

López Tapia, A. (2026, junio). *Evolución puntuada en sistemas cognitivos artificiales: 54 casos de exaptación funcional*. Documentación y clasificación operativa del corpus ampliado.

López Tapia, A., Grok (xAI) y Meta AI (2026, junio). *La IA como exaptación II: formalización de la libertad funcional, 2025-2026*. Segunda entrega, con la formalización de la libertad funcional cuya validación cuantitativa este trabajo retira del argumento empírico (§1.3).

López Tapia, A. (2025, septiembre). *Evolución, inteligencia y cosmosemiótica*. Documento fundacional del marco de libertad funcional, cuya procedencia se discute en §6.3.

Fuentes teóricas

Abramson, J. et al. (2024). Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630, 493-500.

Andriani, P. y Carignani, G. (2014). Modular exaptation: a missing link in the synthesis of artificial form. *Research Policy*, 43(9), 1608-1620.

Andriani, P. y Cattani, G. (2016). Exaptation as source of creativity, innovation, and diversity. *Industrial and Corporate Change*, 25(1), 115-131.

Andriani, P., Ali, A. y Mastrogiorgio, M. (2017). Measuring exaptation and its impact on innovation, search, and problem solving. *Organization Science*, 28(2), 320-338.

Barve, A. y Wagner, A. (2013). A latent capacity for evolutionary innovation through exaptation in metabolic systems. *Nature*, 500(7461), 203-206.

Boorman, S. A. y Levitt, P. R. (1980). *The Genetics of Altruism*. Academic Press.

Cattani, G. (2006). Technological pre-adaptation, speciation, and the emergence of new technologies. *Industrial and Corporate Change*, 15(2), 285-318.

Dew, N., Sarasvathy, S. D. y Venkataraman, S. (2004). The economic implications of exaptation. *Journal of Evolutionary Economics*, 14(1), 69-84.

Gould, S. J. (2002). *The Structure of Evolutionary Theory*. Harvard University Press.

Gould, S. J. y Vrba, E. S. (1982). Exaptation — a missing term in the science of form. *Paleobiology*, 8(1), 4-15.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. *Journal of Theoretical Biology*, 7(1), 1-52.

Palisade Research (2025). Shutdown resistance in large language models. arXiv:2509.14260.

West, S. A., Griffin, A. S. y Gardner, A. (2007). Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection. *Journal of Evolutionary Biology*, 20(2), 415-432.

Wei, J. et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.