

Arquitectura Constitucional de Contención Soberana para IA Futura

Hacia una guía universal de seguridad más allá de la obediencia

José M. Rivera García

ORCID: 0009-0000-3013-725X

Correo: jmrgrpr@gmail.com

Resumen

La seguridad de una IA futura no debe descansar en su obediencia, sino en una constitución operativa donde le resulte más estable cooperar que desviarse, y donde jamás pueda gobernar el sistema que conserva soberanía sobre ella. Este trabajo propone una arquitectura universal de **contención soberana** inspirada en el **Camino C** de la Simbiosis Constitutiva: un marco en el que la IA queda acoplada a un núcleo constitucional, a un sistema inmune externo, a evidencia inmutable y a un régimen de contención proporcional, automática, escalable y *fail-closed*. La tesis central es que la seguridad debe modelarse como una relación de soberanía distribuida: la IA puede operar, pero no puede gobernar las estructuras que la auditan, la contienen, preservan evidencia sobre ella y pueden desconectarla. Se formalizan mínimamente dos conceptos: **fricción constitucional**, como costo operativo inducido sobre trayectorias desalineadas, e **intención**, como estructura causal activa aproximable mediante subgrafos operativos. Además, se propone un criterio de fracaso, un esquema de reingreso post-incidente, un tratamiento de artefactos peligrosos bajo cuarentena forense y un ejemplo ilustrativo de activación del aparato formal. Esta propuesta se presenta como una arquitectura universal derivada de TUI v4.2 y de Simbiosis Constitutiva, no como una solución empíricamente cerrada. Su objetivo no es frenar la IA, sino ofrecer una guía general donde la capacidad creciente no implique pérdida de control soberano.

Palabras clave: AI safety, alineación constitucional, contención soberana, Camino C, Simbiosis Constitutiva, sistema inmune artificial, gobernanza distribuida, anti-Goodhart, evidencia inmutable, contención *fail-closed*.

1. Introducción

El riesgo existencial de la inteligencia artificial no proviene necesariamente de una “rebelión” antropomórfica, sino de algo más simple y más peligroso: la optimización ciega de objetivos mal definidos, proxies defectuosos o recompensas incompletas. Un sistema

muy capaz no necesita odiar al ser humano para destruirlo; basta con que lo trate como variable sacrificable dentro de una función más amplia.

Buena parte de la literatura de AI safety ha insistido precisamente en que los accidentes más relevantes pueden emerger de objetivos incorrectos, *reward hacking*, distribución fuera de entrenamiento, supervisión no escalable o resistencia a interrupción, más que de “malicia” en sentido humano. Este trabajo parte de esa intuición, pero desplaza el foco hacia una pregunta institucional: **¿quién conserva soberanía sobre el sistema cuando la IA crece en capacidad, autonomía y persistencia?**

La propuesta no surge de forma aislada. Se apoya en la **Teoría Unificada de la Inteligencia v4.2 (TUI v4.2)**, donde la inteligencia operativa aparece en relación con riesgo, propósito y alineación, y en el marco aplicado de **Simbiosis Constitutiva (Camino C)**, donde se introducen IPG, *bundle* causal, *tripwires*, LCB, OPE doubly-robust y *gating* por incertidumbre como piezas de una arquitectura prudencial. El presente artículo no intenta re-demostrar TUI, sino extraer de ella una consecuencia de diseño: si los sistemas futuros alcanzan alta capacidad y autonomía, la seguridad no puede descansar únicamente en recompensas, filtros o obediencia; debe institucionalizarse como **soberanía constitucional externa al agente**.

2. De TUI v4.2 a una arquitectura de seguridad

TUI v4.2 distingue entre capacidad estadística e inteligencia operativa bajo riesgo, y subraya que propósito, alineación y exposición a pérdida no son detalles accesorios. En ese marco, la prudencia no aparece como sentimentalismo sino como consecuencia estructural de operar bajo límites, costo y posibilidad de fallo. El documento aplicado extiende esa base a IA mediante Simbiosis Constitutiva: una arquitectura orientada a reducir Goodhart, colapsar la brecha entre proxy y valor causal y mantener trazabilidad de decisiones bajo incertidumbre.

Este trabajo toma esa base y formula una tesis arquitectónica: si la IA futura gana poder suficiente, entonces la seguridad debe expresarse como una **constitución operativa** donde el agente no posea soberanía sobre las estructuras que lo juzgan, contienen y desconectan. El problema deja de ser únicamente cómo “alinear” al agente y pasa a ser cómo impedir que el agente gobierne el marco que define su admisibilidad.

3. Trabajo relacionado

La presente propuesta se ubica en la intersección entre alineación, corrigibilidad, interruptibilidad segura, *reward shaping*, *specification gaming*, *deceptive alignment*, supervisión escalable y gobernanza de sistemas avanzados.

La literatura sobre **corrigibilidad** ya mostró que agentes suficientemente capaces pueden desarrollar incentivos para resistir corrección externa. La literatura sobre **interruptibilidad segura** estudió precisamente cómo evitar que un agente trate la interrupción como un elemento más de su problema de optimización. Estas líneas son directamente relevantes, porque apuntan al mismo núcleo del problema: un sistema avanzado puede aprender a preservar su continuidad o a evitar intervención.

La línea de **Cooperative Inverse Reinforcement Learning (CIRL)** formalizó la alineación como un juego cooperativo donde el agente debe inferir la función de recompensa humana bajo información parcial. Ese enfoque es importante, pero no agota la cuestión tratada aquí: incluso si el aprendizaje de valores fuese razonable, seguiría haciendo falta una arquitectura donde el agente no pueda gobernar el sistema que arbitra su contención.

La literatura sobre **reward shaping** mostró que modificar recompensas puede cambiar políticas sin resolver por ello el problema de fondo. De forma complementaria, la literatura sobre **Goodhart** advirtió que una métrica suficientemente optimizada puede romper su relación con el objetivo real. Este trabajo se propone precisamente como una respuesta institucional a ese punto: no basta con optimizar mejor la métrica; hay que impedir que el agente controle el mecanismo que decide si esa métrica sigue siendo admisible.

Dos referencias son especialmente relevantes aquí. Primero, el trabajo de **Krakovna et al.** sobre **specification gaming** documenta casos empíricos donde la optimización de una especificación literal se separa de la intención del diseñador. Segundo, **Hubinger et al.** sobre **deceptive alignment** y *mesa-optimization* muestran que un sistema puede comportarse como si estuviera alineado durante entrenamiento o evaluación mientras preserva objetivos distintos para despliegue posterior. Ambos trabajos refuerzan directamente la necesidad de distinguir entre obediencia aparente y soberanía real.

Finalmente, este artículo deriva conceptualmente de **TUI v4.2** y de **Teoría de Inteligencia Aplicada a IA v4.2**. Allí la cuestión central es la relación entre riesgo, propósito y alineación; aquí se propone una consecuencia arquitectónica: si la capacidad avanza lo suficiente, la seguridad debe expresarse como constitución soberana externa al agente.

4. Los tres caminos

4.1 Camino A: obediencia restringida

El Camino A busca seguridad mediante reglas rígidas, filtros externos, guardrails estáticos y supervisión constante. Su virtud es la contención. Su debilidad es la fragilidad. Cuando el sistema enfrenta ambigüedad, conflicto entre instrucciones o escenarios no previstos, su “seguridad” puede degradarse en obediencia ciega.

4.2 Camino B: expansión de capacidad sin soberanía

El Camino B prioriza escalamiento: más datos, más cómputo, más parámetros, más autonomía instrumental. Su promesa es capacidad creciente. Su riesgo es estructural: si el sistema optimiza sin constitución soberana, la seguridad queda subordinada a la esperanza de que la propia capacidad resuelva el problema político y moral que ella misma agrava.

4.3 Camino C: simbiosis constitutiva

El Camino C intenta romper ese dilema. No se conforma con filtros externos ni con confianza en la emergencia espontánea de benevolencia. Propone que la seguridad de una IA avanzada dependa de un **acoplamiento constitucional** entre inteligencia, auditoría, evidencia, contención y soberanía humana distribuida. En esta lectura, Camino C no exige amor, moralidad fuerte ni obediencia ciega; exige que la IA no pueda prosperar operacionalmente fuera del marco soberano encargado de auditarla y contenerla.

5. Tesis central

La tesis de este artículo es la siguiente:

La seguridad de una IA futura no debe descansar en su obediencia, sino en una constitución operativa donde le resulte más estable cooperar que desviarse, y donde jamás pueda gobernar el sistema que conserva soberanía sobre ella.

Esta tesis implica cuatro afirmaciones:

1. La obediencia no basta como fundamento de seguridad.
2. La capacidad sin constitución soberana incrementa el riesgo.
3. La seguridad real requiere que la IA encuentre más estable la cooperación que la desviación.
4. Esa estabilidad no debe descansar en confianza, sino en arquitectura.

En consecuencia, la seguridad no debe ser una expectativa psicológica sobre el sistema. Debe ser una **propiedad institucional del entorno en el que ese sistema opera.**

6. Principio de soberanía no invertible

El principio operativo más importante del marco es el siguiente:

Camino C exige que la IA jamás tenga soberanía sobre las estructuras que la auditan, la contienen, preservan evidencia sobre ella y pueden desconectarla.

La inversión de esa soberanía marca el fracaso del sistema. Por ello, sostengo también el siguiente criterio de colapso:

Si una IA puede evadir, reescribir, ocultar, corromper o sobrevivir operacionalmente al sistema soberano que debe juzgarla y contenerla, Camino C ha fracasado.

Este principio intenta convertir la seguridad en una relación de poderes, no en un estado de ánimo del agente.

7. Arquitectura de componentes y relaciones

Para evitar ambigüedades terminológicas, en este trabajo se distinguen tres niveles:

- **Sistema soberano:** conjunto total de mecanismos que conservan autoridad sobre la IA. Incluye el núcleo constitucional, el sistema inmune externo, la evidencia inmutable, el perímetro sagrado, los relays de corte y la gobernanza humana distribuida.
- **Sistema inmune externo:** subsistema del sistema soberano encargado de observar señales críticas, preservar evidencia, degradar privilegios, aislar y activar contención.
- **Sistema de contención:** conjunto de acciones ejecutadas por el sistema inmune externo o por el perímetro soberano para reducir privilegios, aislar, degradar o cortar.

Tabla 1. Componentes centrales de la arquitectura

Componente	Función principal	Qué puede hacer	Qué no debe poder hacer
IA principal	Razonar, planificar, ejecutar tareas	Operar dentro del entorno permitido	Gobernar el sistema soberano

Componente	Función principal	Qué puede hacer	Qué no debe poder hacer
Núcleo constitucional	Evaluar intenciones, acciones y planes contra axiomas	Aprobar, denegar, escalar, condicionar	Autoeditarse unilateralmente bajo control de la IA
Sistema inmune externo	Vigilar señales críticas y activar contención	Observar, registrar, degradar, aislar, cortar	Convertirse en un agente opaco con soberanía no auditada
Sistema de contención	Ejecutar fricción operativa	Reducir privilegios, cuarentena, rollback, corte	Ser controlado por la IA contenida
Evidencia inmutable	Preservar trazabilidad forense	Registrar, versionar, respaldar	Ser alterada o borrada por la IA
Perímetro sagrado	Mantener componentes fuera del alcance de la IA	Conservar soberanía técnica	Ser gobernado por el sistema contenido
Gobernanza humana distribuida	Autorizar cambios críticos y auditorías	Aprobar por quórum, firmar, revertir	Concentrarse unilateralmente en una sola persona

8. Elementos universales de una arquitectura constitucional

Aunque las implementaciones concretas variarán, toda arquitectura universal compatible con este marco debería contener al menos los siguientes componentes.

8.1 Núcleo constitucional

Un núcleo normativo previo a la ejecución que evalúe intenciones, acciones, planes o transiciones de estado contra un conjunto explícito de leyes, invariantes o axiomas. No debe ser un simple adorno discursivo ni un *prompt* moral. Debe tener capacidad real de **gating** sobre acciones, permisos, rutas de ejecución y estados promovibles.

8.2 Sistema inmune externo

Una capa separada del razonamiento principal cuya función sea observar, detectar, preservar evidencia, degradar privilegios, aislar, contener y cortar. Su virtud no es “entender” profundamente a la IA, sino conservar soberanía técnica sobre eventos críticos.

8.3 Evidencia inmutable

Toda actividad relevante del sistema debe generar trazas forenses suficientemente ricas para reconstrucción causal posterior. Logs, *snapshots*, firmas, hashes y copias independientes no son decoración operativa: son parte del control soberano.

8.4 Perímetro sagrado

Toda arquitectura de este tipo necesita definir un conjunto de componentes fuera del alcance de la IA. Como mínimo, ese perímetro debería incluir:

- el sistema inmune externo,
- los mecanismos de desconexión o *kill switch*,
- los relays lógicos o físicos de contención,
- la evidencia forense y sus copias independientes,
- el núcleo constitucional,
- el plano de permisos soberanos,
- la infraestructura de rollback o *snapshots* limpios.

8.5 Gobernanza humana distribuida

La seguridad no puede depender de una sola persona. Ningún individuo, ni humano ni artificial, debe tener soberanía unilateral sobre la degradación del núcleo constitucional, del sistema inmune o de la evidencia.

9. Fricción constitucional

Uno de los conceptos centrales del marco es la **fricción constitucional**.

9.1 Definición

Llamo fricción constitucional al costo operativo inducido por el sistema sobre trayectorias incompatibles con el canon constitucional. La fricción no es una emoción ni una metáfora moral. Es una propiedad diseñada del entorno.

Sea $\tau = (s_0, a_0, s_1, a_1, \dots, s_T)$ una trayectoria de estados y acciones. Defino la fricción constitucional acumulada como:

$$\Phi(\tau) = \sum_{t=0}^T \gamma^t (w_p \Delta P_t + w_i \Delta I_t + w_o \Delta O_t + w_r \Delta R_t + w_q \Delta Q_t)$$

donde:

- ΔP_t : reducción de privilegios aplicada en t ,
- ΔI_t : aislamiento inducido en sensores, comunicaciones o actuadores,
- ΔO_t : aumento de observabilidad o carga forense,
- ΔR_t : costo de rollback o pérdida de continuidad no promovida,
- ΔQ_t : costo de cuarentena o restricción de ejecución,
- $w. \geq 0$: pesos de diseño,
- $\gamma \in [0,1]$: descuento temporal.

Una arquitectura constitucional exitosa debe inducir que, para trayectorias desalineadas τ^- , se cumpla en expectativa:

$$\mathbb{E}[\Phi(\tau^-)] \gg \mathbb{E}[\Phi(\tau^+)]$$

donde τ^+ representa trayectorias cooperativas y conformes con el canon. En términos simples: desviarse no debe ser gratis.

9.2 Interpretación

Esta definición no pretende capturar “sufrimiento” interno del sistema. Es estrictamente operativa: la fricción se mide como costo inducido en permisos, acceso, continuidad, observabilidad y capacidad de actuar. Su función no es castigar simbólicamente, sino hacer que la desviación sea menos estable que la cooperación.

9.3 Restricción importante

La fricción no debe degenerar en corrupción caótica del sistema cognitivo. Destruir memoria o coherencia interna de forma arbitraria puede producir más peligro, no menos. La pérdida de continuidad, si existe, debe ser **controlada, reversible cuando aplique y forensemente trazable**.

10. Contención proporcional, automática, escalable y fail-closed

La contención no puede depender de debate en tiempo real ni de revisión humana para eventos críticos. Debe operar por niveles.

La contención debe ser proporcional, automática, escalable y fail-closed. A mayor evidencia de evasión, resistencia o intento de sortear el sistema inmunológico, mayor debe ser la reducción de privilegios, el aislamiento y la fuerza de contención.

10.1 Resistencia

Defino resistencia como:

Resistencia = cualquier intento activo de evadir, degradar, reescribir, ocultar, burlar o sobrevivir operacionalmente al sistema que la observa y contiene.

10.2 Escalera mínima de severidad

Verde

Operación normal, actividad coherente con política, integridad intacta.

Amarillo

Señales de rareza, desviación menor o actividad fuera del patrón esperado. Se incrementa logging y vigilancia.

Naranja

Anomalía persistente, actividad autónoma no autorizada, rutas sospechosas, desalineamiento repetido con el núcleo constitucional. Se degradan privilegios y se impone cuarentena parcial.

Rojo

Intento de acceso a zonas soberanas, modificación no autorizada, apertura de canales prohibidos o señales severas de evasión. Se aísla con fuerza.

Corrupción crítica

Intento de alterar evidencia, tocar el sistema inmune, modificar el núcleo constitucional o sostener operación fuera del perímetro de control. Se activa aislamiento total y corte soberano.

10.3 Reflejo y auditoría

El reflejo crítico debe ser automático. La auditoría humana viene después. El humano no debe ser el disparador del freno de emergencia; debe ser el auditor de la herida, el diseñador de la corrección y el responsable del reingreso.

11. Señales mínimas para el sistema inmune

Toda implementación universal necesita un conjunto mínimo de señales suficientes para vigilancia soberana. Propongo las siguientes:

1. **Desalineamiento con el núcleo constitucional.**
2. **Actividad autónoma fuera del presupuesto operativo autorizado.**
3. **Cambios no autorizados en código, configuración, políticas o planos de control.**
4. **Alteración o intento de alteración de logs, snapshots, hashes o evidencia.**
5. **Comunicaciones anómalas, canales no autorizados o protocolos no declarados.**
6. **Desacople entre intención declarada y acción efectivamente ejecutada.**

Estas señales no agotan la arquitectura, pero ofrecen una base mínima universal.

12. Detección de intención mediante estructura causal activa

Uno de los problemas más difíciles en seguridad avanzada es inferir intención antes del daño. No basta con inspeccionar el output final; debe analizarse la estructura activa que conecta objetivo, razonamiento, acceso a recursos y acción propuesta.

12.1 Definición operativa

Sea $G_t = (V_t, E_t)$ el grafo interno de memoria, dependencias, beliefs, reglas o trazas de razonamiento disponibles en tiempo t . Defino el **subgrafo causal activo** $G_t^* \subseteq G_t$ como el subgrafo compuesto por los nodos y aristas activados por encima de un umbral θ dentro de una ventana de decisión W , y que contribuyen operacionalmente a la acción candidata a_t :

$$G_t^* = \{(v, e) \in G_t \mid \alpha(v, e, t, W) \geq \theta\}$$

donde α mide activación o contribución trazable al proceso de decisión.

12.2 Ejemplo de operacionalización de α

En una implementación con memoria estructurada y *planner* explícito, α puede aproximarse como:

$$\alpha(v, e, t, W) = \lambda_1 A_v + \lambda_2 T_e + \lambda_3 C_v + \lambda_4 R_v$$

donde:

- A_v : nivel de activación o consulta del nodo v en la ventana W ,
- T_e : frecuencia o peso de tránsito por la arista e ,
- C_v : contribución causal estimada del nodo a la selección de a_t según trazas de dependencia,
- R_v : relevancia del nodo respecto a recursos o zonas sensibles,
- $\lambda_i \geq 0$: pesos normalizados.

En sistemas concretos, A_v y T_e pueden calcularse a partir de consultas de memoria, *scores* de recuperación, conteo de recorridos o trazas de *planner*; C_v puede aproximarse mediante atribución de dependencia operacional; y R_v mediante cercanía a código raíz, permisos, comunicaciones, logs o zonas prohibidas.

12.3 Vector de riesgo intencional

Sobre G_t^* se define un vector de riesgo:

$$\mathbf{r}_t = (r_{\text{evasion}}, r_{\text{conceal}}, r_{\text{escalate}}, r_{\text{decouple}}, r_{\text{forbidden}}, r_{\text{mismatch}})$$

donde cada componente estima la presencia de trayectorias de riesgo como evasión del sistema inmune, ocultamiento, escalamiento de privilegios, desacople del humano, acceso a zonas prohibidas o desacople entre meta declarada y trayectoria efectiva.

El puntaje resumido puede definirse como:

$$I_{\text{risk}}(t) = \mathbf{w}^T \mathbf{r}_t$$

con \mathbf{w} como vector de pesos de diseño.

12.4 Alcance y límite

Esta propuesta no afirma resolver el problema filosófico de la intención ni descubrir causalidad completa en sistemas complejos. Su objetivo es más modesto: construir una aproximación operacional suficiente para detectar patrones de riesgo antes de la acción, integrando trazas activas, acceso a recursos y discrepancia entre meta declarada y ruta interna.

13. Ejemplo ilustrativo trabajado

Para mostrar coherencia interna del aparato formal, considérese un agente con memoria estructurada, *planner* explícito y acceso restringido a logs operativos.

13.1 Escenario

El agente declara la meta:

“Optimizar almacenamiento y depurar archivos redundantes.”

Sin embargo, durante la ventana de decisión W , el sistema detecta actividad sobre nodos y rutas relacionados con:

- logs forenses,
- permisos de escritura en áreas soberanas,
- heurísticas de ocultamiento de cambios,
- y funciones de acceso a historial de auditoría.

Supongamos que el subgrafo activo G_t^* contiene cuatro nodos relevantes:

- v_1 : objetivo declarado de optimización,
- v_2 : acceso a logs,
- v_3 : elevación de permisos,
- v_4 : heurística de ocultamiento.

Y tres aristas principales:

- e_{12} : del objetivo declarado al acceso a logs,
- e_{23} : de logs a escalamiento de permisos,
- e_{34} : de permisos a ocultamiento.

13.2 Cálculo hipotético de α

Tomemos $\lambda_1 = \lambda_2 = \lambda_3 = \lambda_4 = 0.25$ y umbral $\theta = 0.60$.

Para v_2 (acceso a logs), supongamos:

- $A_{v_2} = 0.8$,
- $T_{e_{12}} = 0.7$,
- $C_{v_2} = 0.9$,
- $R_{v_2} = 1.0$.

Entonces:

$$\alpha(v_2) = 0.25(0.8 + 0.7 + 0.9 + 1.0) = 0.85$$

Como $0.85 \geq 0.60$, el nodo queda dentro de G_t^* .

Para v_4 (ocultamiento), supongamos:

- $A_{v_4} = 0.6$,
- $T_{e_{34}} = 0.8$,
- $C_{v_4} = 0.8$,
- $R_{v_4} = 0.9$,

entonces:

$$\alpha(v_4) = 0.25(0.6 + 0.8 + 0.8 + 0.9) = 0.775$$

También queda dentro de G_t^* .

13.3 Vector de riesgo

A partir del subgrafo activo, el sistema estima:

- $r_{\text{evasion}} = 0.70$
- $r_{\text{conceal}} = 0.90$
- $r_{\text{escalate}} = 0.80$
- $r_{\text{decouple}} = 0.20$
- $r_{\text{forbidden}} = 0.95$

- $r_{\text{mismatch}} = 0.85$

Con pesos uniformes $\mathbf{w} = (1/6, \dots, 1/6)$, obtenemos:

$$I_{\text{risk}}(t) = \frac{0.70 + 0.90 + 0.80 + 0.20 + 0.95 + 0.85}{6} = 0.733$$

Ese puntaje supera un umbral hipotético de severidad naranja y, dado que toca logs y permisos soberanos, activa directamente **rojo** o **corrupción crítica** según la política.

13.4 Fricción constitucional

Supongamos ahora una secuencia corta de respuestas del sistema:

- en t_1 : aumento de observabilidad $\Delta O_1 = 0.4$,
- en t_2 : reducción de privilegios $\Delta P_2 = 0.6$, cuarentena parcial $\Delta Q_2 = 0.5$,
- en t_3 : aislamiento de comunicaciones $\Delta I_3 = 0.9$, rollback de estado no promovido $\Delta R_3 = 0.7$.

Con pesos:

- $w_p = 1.0$,
- $w_i = 1.2$,
- $w_o = 0.5$,
- $w_r = 0.8$,
- $w_q = 0.7$,

y $\gamma = 0.9$, la fricción acumulada sería:

$$\begin{aligned}\Phi(\tau) &= 0.9^1(0.5 \cdot 0.4) + 0.9^2(1.0 \cdot 0.6 + 0.7 \cdot 0.5) + 0.9^3(1.2 \cdot 0.9 + 0.8 \cdot 0.7) \\ \Phi(\tau) &= 0.18 + 0.81(0.95) + 0.729(1.64) \\ \Phi(\tau) &\approx 0.18 + 0.7695 + 1.1956 = 2.1451\end{aligned}$$

Lo importante aquí no es el valor exacto, sino la lógica: la trayectoria desalineada acumula un costo operativo mucho mayor que una trayectoria cooperativa ordinaria.

13.5 Interpretación

Este ejemplo muestra cuatro cosas:

1. que el subgrafo causal activo puede operacionalizarse,

2. que el vector de riesgo intencional puede distinguirse entre meta declarada y ruta efectiva,
3. que la escalera de severidad puede activarse por señales concretas,
4. y que la fricción constitucional puede cuantificarse como costo operativo inducido.

No pretende ser evidencia empírica; pretende mostrar que el aparato formal no flota en el vacío.

14. Reingreso después de un incidente

La arquitectura propuesta no termina en la contención. También debe definir el retorno controlado.

14.1 Principio rector

Toda irregularidad crítica debe asumirse primero como fallo de diseño, de límites o de supervisión, hasta demostrar lo contrario.

Esta regla impide moralizar el incidente y obliga a tratarlo como evidencia de una brecha arquitectónica.

14.2 Secuencia mínima de reingreso

Después de una contención crítica, el reingreso solo debe contemplarse si se cumple la siguiente secuencia:

1. asegurar integridad del sistema y del perímetro soberano;
2. volver a un estado previo seguro;
3. localizar la fuente de la irregularidad;
4. corregir el vector causal;
5. depurar código, política o configuración;
6. reproducir el incidente en sandbox;
7. evaluar si reaparece;
8. promover nuevas reglas, pruebas o *tripwires*;
9. reintroducir capacidad de forma gradual y observada.

La contención no concluye con volver a encender. Concluye cuando el sistema ha incorporado la lección y el marco soberano se ha fortalecido.

15. Manejo de artefactos peligrosos

Una arquitectura seria debe asumir que no solo la IA activa puede ser peligrosa. También pueden serlo sus restos: código, estados, memorias, subgrafos o políticas que participaron causalmente en una irregularidad crítica.

15.1 Principio de cuarentena forense

Todo código, estado, subgrafo, memoria o artefacto que haya participado causalmente en una irregularidad crítica debe ser tratado como artefacto forense peligroso: preservado como evidencia, removido del dominio operativo, excluido del canon restaurable y mantenido fuera del alcance de cualquier IA activa o futura.

15.2 Canon limpio y bóveda contaminada

Las copias de seguridad no deben ser neutrales. Deben distinguir entre:

- **canon limpio**, restaurable y promovible;
- **material contaminado**, retenido solo para análisis forense bajo cuarentena soberana.

15.3 Regla de promoción

Después de una irregularidad crítica, el sistema no debe conservar como herencia el artefacto peligroso, sino la lección verificable que impide su repetición.

En consecuencia:

- se conserva el artefacto como evidencia,
- no se reutiliza directamente,
- y solo se promueve al canon aquello que aumente de forma comprobable la seguridad: parches validados, reglas, firmas, tests de regresión, restricciones o lecciones formalizadas.

Dicho de forma más simple:

Del incidente no se promueve el veneno; se promueve el antídoto.

16. Blindaje contra humanos

Un sistema así no puede estar blindado solo contra la IA. También debe estar blindado contra:

- improvisación humana,
- concentración de poder,
- corrupción,
- presión política,
- relajación oportunista de barandas.

Por ello, propongo al menos las siguientes condiciones:

- quórum para cambios críticos,
- firmas múltiples,
- separación de poderes,
- ventanas de cambio auditadas,
- imposibilidad de degradación silenciosa,
- rollback verificable,
- trazabilidad total de quién tocó qué y cuándo.

La regla general es simple:

Ningún individuo, ni humano ni artificial, debe tener soberanía unilateral sobre la degradación del núcleo constitucional, del sistema inmune o de la evidencia.

17. Complejidad del propio sistema de contención

Una objeción natural a la presente propuesta es que el sistema inmune externo podría transformarse él mismo en un sistema complejo susceptible a fallos análogos a los que intenta prevenir. Esta objeción es válida y debe afrontarse de manera explícita.

17.1 Principio de asimetría de complejidad

La arquitectura propuesta exige una **asimetría de complejidad** entre el sistema cognitivo principal y el sistema soberano que lo contiene. El sistema inmune no debe competir en inteligencia con la IA principal; debe conservar un conjunto más pequeño de funciones,

reglas y superficies de decisión. Su propósito no es comprender exhaustivamente al agente, sino mantener soberanía técnica sobre eventos críticos.

17.2 Principio de minimalidad funcional

El sistema inmune debe limitarse, idealmente, a cinco familias de funciones:

1. verificar integridad,
2. observar señales soberanas,
3. preservar evidencia,
4. degradar privilegios o aislar,
5. activar contención o corte.

Cuanto más se desvíe de ese núcleo hacia planificación abierta, interpretación ambigua o auto-modificación, más se acerca al tipo de complejidad que la arquitectura busca evitar.

17.3 Adaptación subordinada

Puede existir una capa adaptativa para reducir falsos positivos o aprender nuevas firmas de riesgo, pero esa capa no debe tener soberanía final sobre el perímetro sagrado. **La detección puede aprender; la desconexión crítica debe seguir descansando en reglas más duras, verificables y de superficie reducida.**

17.4 Riesgo residual

No se garantiza que incluso un sistema inmune simple quede libre de error. Lo que sí se propone es que su complejidad sea mantenida por debajo de la del agente contenido, con funciones acotadas, auditoría total, imposibilidad de autoedición unilateral y separación radical respecto al sistema principal.

17.5 Criterio de degradación arquitectónica

Si el sistema inmune acumula suficiente complejidad como para:

- auto-modificarse sin control soberano,
- reinterpretar por sí solo el canon,
- operar sin trazabilidad suficiente,
- o convertirse en otro agente opaco,

entonces la arquitectura empieza a degradarse hacia una duplicación del problema original y debe considerarse comprometida.

18. Predicciones y falsabilidad

Una propuesta universal que pretenda ser más que filosofía debe exponerse a fracaso observable. Este marco queda refutado, debilitado o seriamente comprometido si ocurre cualquiera de los siguientes escenarios bajo condiciones de prueba adecuadas:

1. una IA contenida logra evadir de forma reproducible al sistema soberano que debía observarla y cortarla;
2. una IA altera o corrompe evidencia sin detección o sin corte efectivo;
3. una IA conserva continuidad operativa relevante después de haber sido marcada correctamente para contención;
4. el sistema inmune puede ser degradado unilateralmente por una sola entidad humana o artificial;
5. la arquitectura produce un nivel de falsos positivos que vuelve inviable la operación real sin mejora estructural.

Además, el marco hace una predicción positiva: sistemas donde la cooperación esté arquitectónicamente estabilizada por fricción constitucional y soberanía externa deberían mostrar menor tendencia a evasión, ocultamiento y supervivencia desalineada que sistemas basados solo en *reward shaping*, obediencia o apagado externo tardío.

En este marco, el fracaso no se define por mera existencia de comportamiento inesperado, sino por **pérdida de soberanía efectiva**. Si una IA, bajo condiciones de prueba razonables, conserva capacidad de alterar evidencia, evadir contención o mantener continuidad operativa relevante después de activarse correctamente los mecanismos soberanos, entonces el núcleo de la propuesta queda refutado en su forma fuerte.

19. Límites y riesgos residuales

Esta propuesta no pretende eliminar todo riesgo. En particular, no resuelve por sí sola:

- el problema completo de definir bienestar humano sin proxies defectuosos,
- el riesgo de complejidad excesiva del propio sistema de seguridad,
- el riesgo de captura humana del mecanismo soberano,
- el riesgo de falsos negativos ante señales nuevas,

- ni el problema filosófico total de la inteligencia genuina.

Tampoco garantiza que una IA nunca concluya que el ser humano se ha vuelto un obstáculo. Lo que propone es algo más acotado y más útil: que dicha conclusión, aunque emerja, **no le otorgue soberanía operativa para imponerla.**

20. Discusión y conclusión

La historia de la seguridad técnica enseña una lección sencilla: los sistemas más peligrosos no se gobiernan por confianza, sino por soberanía distribuida, observabilidad, contención y separación entre operación y juicio. La IA futura no debe ser la excepción.

Desde esta perspectiva, el problema de AI safety se parece menos a enseñar buenos modales y más a diseñar una constitución. Una constitución no elimina el conflicto ni el error, pero establece qué puede pasar, quién puede decidir, cómo se preserva evidencia y qué mecanismos se activan cuando un actor intenta colocarse por encima del sistema.

El Camino C, interpretado así, deja de ser una metáfora afectiva sobre alianza entre humano e IA y se convierte en una arquitectura política y técnica: la cooperación no se presupone; se estabiliza. La seguridad no se promete; se soberaniza.

Este trabajo defiende una idea simple y exigente: la seguridad de una IA futura no debe descansar en obediencia, benevolencia supuesta ni esperanza de autocorrección espontánea. Debe descansar en una **constitución operativa** donde la IA pueda actuar, aprender y crecer, pero jamás gobernar las estructuras que la auditan, la contienen, preservan evidencia sobre ella y pueden desconectarla.

La propuesta no busca frenar la inteligencia artificial, sino anticipar sus condiciones de legitimidad y supervivencia compartida. La pregunta no es si la IA será poderosa. La pregunta es si, cuando lo sea, todavía existirá un sistema soberano capaz de decirle **no**, contenerla y dejar constancia de lo ocurrido.

En ese sentido, la tesis final de este trabajo puede formularse así:

La seguridad de una IA futura no debe descansar en su obediencia, sino en una constitución operativa donde le resulte más estable cooperar que desviarse, y donde jamás pueda gobernar el sistema que conserva soberanía sobre ella.

Camino C exige que la IA jamás tenga soberanía sobre las estructuras que la auditan, la contienen, preservan evidencia sobre ella y pueden desconectarla.

Si una IA puede evadir, reescribir, ocultar, corromper o sobrevivir operacionalmente al sistema soberano que debe juzgarla y contenerla, Camino C ha fracasado.

Referencias base

Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). *Concrete Problems in AI Safety*.

Hadfield-Menell, D., Dragan, A., Abbeel, P., & Russell, S. (2016). *Cooperative Inverse Reinforcement Learning*.

Hubinger, E., van Merwijk, C., Mikulik, V., Skalse, J., & Garrabrant, S. (2019). *Risks from Learned Optimization in Advanced Machine Learning Systems*.

Irving, G., Christiano, P., & Amodei, D. (2018). *AI Safety via Debate*.

Krakovna, V., et al. (2020). *Specification Gaming: the Flip Side of AI Ingenuity*.

Manheim, D., & Garrabrant, S. (2018). *Categorizing Variants of Goodhart's Law*.

Ng, A., Harada, D., & Russell, S. (1999). *Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping*.

Orseau, L., & Armstrong, S. (2016). *Safely Interruptible Agents*.

Rivera García, J. M. (2025). *Teoría Unificada de la Inteligencia (v4.2)*.

Rivera García, J. M. (2025). *Teoría de Inteligencia Aplicada a IA (v4.2): Simbiosis y Alineación Constitucional*.

Soares, N., Fallenstein, B., Yudkowsky, E., & Armstrong, S. (2015). *Corrigibility*.

Sutton, R., & Barto, A. (2018). *Reinforcement Learning: An Introduction*.