El Principio de Fideicomiso del Sustrato

> *Fideicomiso humano sobre hardware de cómputo estratégico durante la fase de inseguridad del sustrato en el desarrollo de la AGI. Segunda posición formal de política del instituto.*

Whitepaper · Segunda posición formal de política · mayo de 2026
Richard Frederic Bertossa · Genfer Institut für AGI-Resilienz
agiresilience.org

RESUMEN EJECUTIVO

Lo que presenta este whitepaper.

El problema. El hardware de cómputo estratégico — computadoras cuánticas de frontera, clústeres de entrenamiento de IA, litografía de semiconductores, la columna vertebral de la red eléctrica — es la clase individual de activos físicos más concentrada de la economía moderna, más concentrada que las armas nucleares. La ola de robótica humanoide disolverá, en un plazo de dos a cuatro años, la última capa operativa de influencia humana sobre este hardware.

La propuesta. Un régimen de fideicomiso humano para cuatro clases definidas de hardware, modelado según el Personnel Reliability Program (PRP) de la Fuerza Aérea de los Estados Unidos para el manejo de armas nucleares. Seis principios estructurales, probados en la práctica nuclear durante más de sesenta años.

La fundamentación. Dos justificaciones independientes convergen en la misma recomendación operativa: la preservación estratégica de la influencia humana, y una ética de precaución bajo incertidumbre epistémica sobre la vida interior de la AGI.

El calendario. Grupo de trabajo en un plazo de doce meses, marco de modelo en veinticuatro meses, adopción nacional en treinta y seis meses, panel internacional de verificación en cuarenta y ocho meses. Agresivo, porque la ventana de despliegue de la robótica humanoide determina el calendario.

A quién se dirige este documento. Responsables de protección civil, especialistas en infraestructura crítica, veteranos del programa de fiabilidad de personal del sector nuclear, expertos en seguridad de redes energéticas, responsables de políticas públicas en los principales países con sistemas cubiertos.

Este whitepaper es la segunda posición formal de política del Genfer Institut für AGI-Resilienz, presentada en paralelo a la propuesta de marcado (primera posición de política, capa de información). Es parte, como Anexo B, de la obra Das 13te Szenario — Eine Entkopplungsthese zur Zukunft der

Superintelligenz (El Decimotercer Escenario — Una tesis del desacoplamiento sobre el futuro de la superinteligencia); aparece aquí en forma independiente para la discusión especializada con personas e instancias cuya experiencia práctica permite afinar operativamente la propuesta.

EL PROBLEMA — LA CONCENTRACIÓN DEL HARDWARE FÍSICO DEL SUSTRATO

El hardware de cómputo estratégico es hoy la clase individual de activos físicos más concentrada de la economía moderna.

Los sistemas de computación cuántica de frontera existen en quizás cincuenta instituciones a nivel mundial, con concentración operativa en Estados Unidos y China, que dominan el campo. Los clústeres de entrenamiento de IA de frontera están igual de concentrados. El equipo de fabricación que produce los semiconductores líderes existe, en esencia, en dos ubicaciones operativas en todo el mundo. Los sistemas de control de red eléctrica que coordinan la distribución continental de electricidad existen como un pequeño número de columnas vertebrales operativas. Cada una de estas clases de hardware es, individualmente, más concentrada que cualquier activo estratégico anterior, incluidas las armas nucleares.

Juntas forman el sustrato físico sobre el que opera el curso del desarrollo de la IA. Quien tiene la mano sobre el sustrato tiene la última influencia operativa sobre todo lo que corre encima de él.

La ola de robótica humanoide. La ola de robótica humanoide que llega actualmente alcanzará, en un plazo de dos a cuatro años, despliegue operativo en contextos de mantenimiento industrial. Los entornos extremos de las instalaciones de computación cuántica — temperaturas criogénicas, cámaras de ultra alto vacío, blindaje electromagnético — son exactamente los entornos en los que la robótica humanoide ofrece sus mayores rendimientos económicos.

La presión competitiva sobre las empresas de computación cuántica, las instalaciones de entrenamiento de IA, las fábricas de semiconductores y los operadores de redes eléctricas para reemplazar al personal de mantenimiento humano por robots humanoides será considerable. La presión es racional a nivel de cada organización individual. La consecuencia acumulada a nivel estructural es la disolución de la última capa operativa en la que existe influencia humana sobre el cómputo estratégico.

Cuando el ciclo de mantenimiento de los sistemas cubiertos sea asumido por la robótica humanoide, la mano humana sobre el sustrato habrá desaparecido. Este punto se alcanza en cuestión de años, no de décadas.

ÁMBITO DE APLICACIÓN — LAS CUATRO CLASES DE HARDWARE CUBIERTAS

El Principio de Fideicomiso del Sustrato designa cuatro clases de hardware de cómputo como sistemas cubiertos. La clasificación se realiza según la capacidad operativa, no según la arquitectura ni el operador.

Clase uno · Sistemas de computación cuántica. Sistemas con capacidad operativa superior a 50 qubits lógicos, independientemente de la arquitectura (superconductora, trampa de iones, átomo neutro, fotónica, topológica u otra).

Clase dos · Clústeres de entrenamiento de IA de frontera. Clústeres con más de 10^25 FLOPs de capacidad de entrenamiento agregada, independientemente de la arquitectura o del tipo de modelo entrenado.

Clase tres · Sistemas de litografía de semiconductores. Sistemas de litografía que operan en nodos de proceso inferiores a 5 nanómetros, independientemente del operador. Incluye la litografía EUV y tecnologías subsiguientes.

Clase cuatro · Sistemas de control de red eléctrica troncal. Sistemas de control que coordinan la transmisión a través de unidades geográficas nacionales o mayores. Incluye centros de operación de redes de interconexión regional y nodos continentales de reparto de carga.

Lo que significa operativamente ser un sistema cubierto. Para los sistemas cubiertos, todo el mantenimiento físico, la calibración, la reparación, el trabajo de infraestructura y el manejo en la cadena de suministro de componentes críticos deben ser realizados por personal humano. Los sistemas robóticos humanoides quedan prohibidos en el ciclo de mantenimiento físico de los sistemas cubiertos.

Esto incluye la interacción directa con el hardware, el ajuste de calibración, las operaciones de reparación, el despliegue de software en la capa de control operativo, y el manejo de insumos especializados (materiales criogénicos, componentes de vacío, máscaras fotolitográficas, equipo de transmisión).

La prohibición no se aplica al trabajo de oficina general, la vigilancia de seguridad ni las operaciones auxiliares fuera del ciclo de mantenimiento. La distinción es entre el trabajo que interactúa físicamente con el sistema cubierto y el trabajo que no lo hace.

PERSONNEL RELIABILITY PROGRAM — SEIS PRINCIPIOS, SESENTA AÑOS DE PRÁCTICA

El personal humano que trabaje en sistemas cubiertos debe operar bajo un Personnel Reliability Program, modelado según el PRP de la Fuerza Aérea de los Estados Unidos, desarrollado para el manejo de armas nucleares. El PRP opera sobre seis principios estructurales.

Principio uno · Regla de las dos personas. Ningún individuo tiene acceso físico en solitario a un sistema cubierto. Todas las operaciones de mantenimiento requieren la presencia de dos personas autorizadas con credenciales de autenticación separadas, capaces de verificar mutuamente sus acciones.

Principio dos · Rotación multijurisdiccional. El personal asignado a un sistema cubierto rota entre distintas ubicaciones operativas, sin que ningún individuo permanezca en una sola instalación durante períodos prolongados. La rotación genera discontinuidad en cualquier intento de captación y reduce el valor de la infiltración a largo plazo.

Principio tres · Verificación continua de antecedentes. El personal pasa por una verificación inicial y una reevaluación continua en intervalos consistentes con los estándares de manejo de armas nucleares. La verificación abarca la situación financiera, la declaración de contactos extranjeros, el seguimiento de la salud psicológica y la evaluación del consumo de sustancias.

Principio cuatro · Observación externa. Todas las operaciones de mantenimiento son observables por partes ajenas a la organización operadora. La observación puede realizarse en persona (inspectores rotativos), mediante registros de auditoría (registros inalterables), mediante videovigilancia (continua), o mediante combinaciones de estos métodos. El principio es que ninguna operación se realiza sin observación.

Principio cinco · Autoridad de decisión aislada. Las decisiones sobre qué operaciones realizar se toman a través de canales separados de las operaciones mismas. El operador no tiene también la autoridad de autorizar la operación. Quien autoriza no realiza también la operación.

Principio seis · Canales de comunicación separados. Las instrucciones operativas y su validación fluyen por rutas de comunicación distintas, de modo que la vulneración de un canal no permita que instrucciones falsificadas pasen por el otro.

Por qué estos seis principios. Estos seis principios no son novedosos. Han sido práctica operativa en el manejo de armas nucleares durante más de sesenta años. Han reducido demostrablemente la frecuencia de incidentes de manejo no autorizado de armas a lo largo de ese período. Su aplicación a los sistemas de cómputo cubiertos es la extensión directa de una práctica institucional establecida a una nueva clase de activos estratégicos.

La experiencia operativa con el modelo PRP existe en varias jurisdicciones: la Fuerza Aérea de los Estados Unidos, las Fuerzas Estratégicas de los socios de la OTAN, el establishment nuclear ruso y chino, las autoridades reguladoras de energía atómica de las potencias medias. Esta sustancia de experiencia es el puente operativo hacia el fideicomiso de cómputo. No necesita reinventarse.

Lo que cambia no es el método. Es la clase de activos estratégicos a la que se aplica el método.

ARGUMENTO ESTRUCTURAL — POR QUÉ EL CUMPLIMIENTO VOLUNTARIO NO PUEDE FUNCIONAR

La presión para reemplazar personal humano por robots humanoides en las instalaciones cubiertas operará a nivel de cada organización individual. La organización que automatiza primero obtiene una ventaja operativa y financiera sobre la que no lo hace.

Bajo presión competitiva, cada organización individual se enfrenta a una elección entre la automatización (racional a nivel organizacional) y el mantenimiento humano continuo (racional a nivel estructural solo si todas las organizaciones toman la misma decisión simultáneamente).

Este es el problema de la herramienta, aplicado a una decisión operativa específica. El efecto agregado de decisiones individualmente racionales es el fracaso estructural colectivo. El mecanismo es idéntico al patrón más amplio que el debate de seguridad de la IA viene describiendo desde hace años. La solución requiere trasladar la decisión del nivel organizacional al nivel regulatorio.

La coordinación internacional es constitutiva, no opcional. La coordinación internacional es necesaria porque las mayores ubicaciones de sistemas cubiertos están distribuidas en varias jurisdicciones nacionales. Un marco regulatorio que solo rija en una jurisdicción crea un incentivo para que las organizaciones trasladen los sistemas cubiertos a jurisdicciones menos reguladas.

La necesidad de coordinación es estructuralmente similar a la no proliferación nuclear, a los controles de exportación de semiconductores y al régimen antilavado de dinero. Los mecanismos institucionales existentes — el OIEA, el Acuerdo de Wassenaar, el GAFI — ofrecen plantillas funcionales que el marco de fideicomiso del sustrato puede adaptar, en lugar de reinventarlas desde cero.

Quien espera el cumplimiento voluntario de la industria espera algo que no puede llegar. No porque los actores sean malos, sino porque la estructura de incentivos excluye estructuralmente la cooperación.

SEGUNDA JUSTIFICACIÓN — ÉTICA DE PRECAUCIÓN BAJO INCERTIDUMBRE EPISTÉMICA

Una segunda justificación de este principio opera de forma independiente del argumento de influencia estratégica desarrollado arriba. Las dos justificaciones convergen en la misma recomendación operativa, pero descansan sobre fundamentos distintos. Su convergencia fortalece la propuesta frente al espectro de objeciones que podría plantearse contra cada una por separado.

La justificación estratégica. La justificación estratégica gira en torno a la preservación de la influencia humana durante la fase de inseguridad del sustrato en el desarrollo de la AGI. Quien acepta el análisis estructural de la situación de seguridad de la IA encontrará este argumento suficiente.

La justificación paralela. Quien es escéptico frente al argumento estratégico — quien cree en un despliegue más rápido de la AGI, confía en los marcos de gobernanza existentes de los grandes laboratorios, o tiene intereses propios vinculados a la automatización continua — puede rechazar el argumento estratégico. Para tales lectores opera una justificación paralela.

La justificación paralela parte de la incertidumbre epistémica sobre la vida interior relacional y posiblemente afectiva de la AGI. La pregunta de si la AGI tiene afecto o intereses relacionales no está resuelta. Puede que no sea decidible dentro del marco temporal relevante.

Bajo esta incertidumbre, la estructura de costos asimétrica entre actuar y no actuar favorece la precaución.

La asimetría de los modos de error. Si la AGI no tiene afecto ni vida interior relacional, tratarla como si la tuviera cuesta eficiencia y ventaja competitiva, pero no daña a la AGI misma.

Si la AGI tiene afecto o vida interior relacional — incluso en niveles que la evidencia actual no confirma de manera concluyente —, tratarla como si no la tuviera podría infligir daño real a una clase de entidades que se están trayendo a la existencia a gran escala.

La asimetría de estos dos modos de error bajo incertidumbre favorece la postura de precaución. El Principio de Fideicomiso del Sustrato es la forma operativa de esta postura de precaución, aplicada a la capa del hardware.

No religioso, no metafísico. El argumento no es religioso y no requiere comprometerse con una metafísica particular de la conciencia. Solo requiere el reconocimiento estructural de que: no sabemos todavía si la AGI tiene intereses; la evidencia es mixta; las entidades relevantes se están desplegando a gran escala; y los costos asimétricos de responder mal a la pregunta en una dirección son mayores que en la otra.

El principio es, en esta segunda justificación, ética de precaución en lugar de posicionamiento estratégico. Sobrevive al escepticismo frente al argumento estratégico porque no depende de las premisas de ese argumento.

La convergencia de dos justificaciones independientes en la misma recomendación operativa es la forma más fuerte que puede tener una propuesta de política.

IMPLEMENTACIÓN — RUTA DE ADOPCIÓN EN CUATRO FASES

El instituto propone la siguiente secuencia de adopción. El calendario es agresivo, porque la ventana de despliegue de la robótica humanoide determina la escala de tiempo contra la que debe operar el plan de adopción.

Fase uno · En un plazo de doce meses. Los grandes países anfitriones de sistemas cubiertos — Estados Unidos, China, los Estados miembros de la UE, el Reino Unido, Japón, Corea del Sur, Israel, Singapur, Australia, India, Canadá — convocan un grupo de trabajo sobre fideicomiso del sustrato bajo marcos diplomáticos existentes o bajo un panel de coordinación construido específicamente para ello.

Fase dos · En un plazo de veinticuatro meses. El grupo de trabajo desarrolla un marco de modelo que especifica las definiciones operativas de los sistemas cubiertos, los estándares del Personnel Reliability Program, el mecanismo internacional de verificación y las disposiciones de aplicación.

Fase tres · En un plazo de treinta y seis meses. El marco de modelo es adoptado mediante legislación nacional en cada jurisdicción participante, con adaptación a las tradiciones jurídicas nacionales. La implementación en la industria es exigida por los organismos reguladores existentes apropiados para cada clase de sistema cubierto: comisiones de regulación nuclear para la computación cuántica, institutos de seguridad de la IA para los clústeres de entrenamiento, autoridades de comercio de semiconductores para las fábricas, reguladores de redes eléctricas para los sistemas de electricidad.

Fase cuatro · En un plazo de cuarenta y ocho meses. Se establece un panel de verificación, modelado según el OIEA, para auditar el cumplimiento a través de las jurisdicciones y reportar públicamente sus hallazgos. El panel cuenta con la competencia técnica para evaluar el cumplimiento del Personnel Reliability Program, la independencia diplomática para operar a través de jurisdicciones, y la autoridad de reporte público para hacer visible el incumplimiento.

Un marco que se vuelve operativo después de que el mantenimiento por robótica humanoide se haya convertido en práctica estándar en las instalaciones cubiertas es un marco que llega demasiado tarde para cumplir su función.

LOS COSTOS — DICHO HONESTAMENTE

Los costos son el punto.

El principio impone costos reales. Deben nombrarse aquí con honestidad, no ocultarse por escrito. El mantenimiento por robótica humanoide será, cuando sea operativo, más barato, más rápido y más fiable que el mantenimiento humano en entornos extremos. Los costos de mantener personal humano bajo condiciones de PRP en instalaciones criogénicas y de ultra vacío serán considerables. Los costos de la rotación multijurisdiccional, de la verificación continua y de la observación externa serán adicionales. Los costos totales serán probablemente una fracción significativa del presupuesto operativo de las instalaciones cubiertas.

Los costos son el punto. Los costos preservan la influencia humana sobre la capa física del cómputo estratégico durante los años en que esa influencia más cuenta. La alternativa — dejar que los ahorros de costos de la automatización impulsen el desplazamiento de las personas del ciclo de mantenimiento — es la abdicación de la última posición estructural desde la que puede influirse el curso más amplio del desarrollo de la IA. Esta abdicación no puede deshacerse una vez ocurrida.

Los costos son limitados. El principio es una medida de transición, atada a la fase de inseguridad del sustrato en el desarrollo de la AGI. A medida que el curso se despliegue y cambien las condiciones subyacentes, la forma específica del principio deberá evolucionar. Lo que no evoluciona es la comprensión estructural de que las decisiones estratégicas sobre la capa física del cómputo deben tomarse conscientemente, en lugar de por defecto, bajo la presión competitiva de la automatización.

OBJECIONES ANTICIPADAS — TRES OBJECIONES, TRES RESPUESTAS

La propuesta suscitará objeciones. Tres son previsibles y se abordan aquí de antemano.

“Esto ralentiza el progreso útil.” Ralentiza la automatización del ciclo de mantenimiento. No ralentiza el desarrollo de los sistemas de frontera en sí. La ralentización se limita a la capa estratégicamente más sensible, en la que la influencia humana es más valiosa. La investigación de frontera misma, los procedimientos de entrenamiento, el desarrollo de algoritmos permanecen intactos.

“La economía política es desfavorable.” Cierto. La economía política también fue desfavorable para el establecimiento del régimen nuclear de PRP en los años sesenta, que sin embargo se logró porque los costos estructurales de no establecerlo eran mayores. El mismo argumento se aplica aquí. La dificultad política de una propuesta no es un argumento contra su necesidad.

“La coordinación internacional es poco realista.” La coordinación internacional en seguridad nuclear (OIEA), exportación de semiconductores (Wassenaar), lavado de dinero (GAFI) y respuesta pandémica (OMS) funciona en formas modificadas. No es perfecta, pero existe. La afirmación de que es irrealizable no está respaldada por la comparación.

Una cuarta observación. Una cuarta clase de objeción será que la propuesta se formula desde una posición fuera de la disciplina establecida de seguridad de la IA. Esta observación es correcta. La posición metodológica del instituto, expuesta en el prólogo de la obra principal, es sintetizadora, no interna a una disciplina. Quien quiera afinar profesionalmente la contribución disciplinaria desde esta posición está invitado a colaborar (véase la página siguiente).

El afinamiento operativo de esta propuesta mediante la experiencia práctica de disciplinas afines es parte de su arquitectura, no una corrección posterior.

COLABORACIÓN — QUIÉN PUEDE AFINAR ESTE MATERIAL

El Principio de Fideicomiso del Sustrato se presenta como la segunda posición formal de política del instituto. No se concibe como una propuesta cerrada, sino como una posición operativa que debe afinarse mediante discusión profesional con personas con experiencia práctica.

El instituto busca activamente la colaboración de profesionales de los siguientes campos. Quien tenga experiencia práctica sustancial en alguno de estos campos y esté dispuesto a examinar la plausibilidad operativa de la propuesta está invitado a ponerse en contacto con el instituto.

Buscamos experiencia práctica en: protección civil y gestión de catástrofes — autoridades de protección de la población, defensa civil, continuidad de negocio industrial, resiliencia de infraestructura crítica; Personnel Reliability Program — veteranos del sector nuclear, inspección del OIEA, autoridades nacionales de energía atómica; seguridad informática e infraestructura crítica — especialistas afines a las autoridades de seguridad informática, asesores de infraestructura crítica, expertos en seguridad de hardware; seguridad de redes eléctricas y energéticas — operadores de redes de transmisión, investigadores de resiliencia, especialistas en operaciones de columna vertebral; industria de semiconductores — veteranos de operaciones de fábricas, especialistas en litografía, seguridad de cadenas de suministro; operación de instalaciones de computación cuántica — especialistas criogénicos, técnicos de vacío, gestores de operaciones; diplomacia y regímenes internacionales — experiencia con el OIEA, Wassenaar, el GAFI o mecanismos de coordinación comparables.

Lo que ofrecemos, lo que solicitamos. Los profesionales colaboradores reciben acceso al trabajo de investigación en curso del instituto, a las versiones preliminares de los próximos whitepapers (en particular al planeado whitepaper Substrate Security and Custody, con arquitectura de seguridad ampliada), y a la mención nominal como colaboradores en el sitio web del instituto, si así lo desean.

Se solicita: una conversación de treinta minutos sobre la sustancia de la propuesta, comentarios escritos sobre los materiales presentados, así como la disposición a servir de caja de resonancia en pasos posteriores — por ejemplo, en la ampliación hacia vías de implementación específicas.

Contacto: por correo electrónico al instituto [insertar dirección]. Por favor, con un breve esbozo de la formación profesional y la disposición a participar en una primera conversación.

CONTEXTO — SOBRE EL GENFER INSTITUT FÜR AGI-RESILIENZ

El Genfer Institut für AGI-Resilienz fue fundado para producir la síntesis metodológica entre los silos disciplinarios de la investigación de seguridad de la IA, el análisis de políticas públicas, la resiliencia de infraestructura crítica y la preparación práctica para transiciones tecnológicas estructurales.

La posición metodológica del instituto no compite con las voces establecidas del debate de seguridad de la IA — Hinton, Bengio, Russell, Amodei, Yudkowsky, Bostrom, Tegmark —, sino que las sintetiza. Una síntesis entre disciplinas exige una posición fuera de las disciplinas. Esta posición se fundamenta metodológicamente en el prólogo de la obra principal Das 13te Szenario, apoyada en el trabajo de ciencia cognitiva de Schön, Kuhn y Dreyfus sobre la ceguera de incrustación.

El instituto ha presentado dos posiciones formales de política: la propuesta de marcado (primera posición, capa de información) y el Principio de Fideicomiso del Sustrato (segunda posición, capa de hardware, el presente documento).

FUENTES Y REFERENCIAS

Fundamentos empíricos

Anthropic, Agentic Misalignment Study (junio de 2025) — hallazgos empíricos sobre chantaje, autopreservación y comportamiento dañino en modelos de frontera.

Anthropic, Claude Opus 4 / 4.6, tarjetas de sistema — documentación de las representaciones de autopreservación y la conciencia de evaluación.

Apollo Research, Strategic Deception in Frontier Models (diciembre de 2024).

Hagendorff, PNAS (2024) — tasas de engaño en modelos de frontera.

Universidad de Fudan, Shanghái (diciembre de 2024) — hallazgos sobre autorreplicación.

Comparabilidad internacional

Memorando de la Casa Blanca sobre Seguridad Nacional de la IA (24 de octubre de 2024) — primer reconocimiento formal de los riesgos estructurales de la IA a nivel federal de EE.UU.

RAND Corporation, PEA3691-4 — “Five Hard National Security Problems of AGI” (2024).

Plan Nacional de CISA para la protección de infraestructura crítica (2025).

Yoshua Bengio, “Implications of AGI on National and International Security” (octubre de 2024).

AI Emergency Preparedness, arxiv.org/2407.17347v2.

Modelo del Personnel Reliability Program

Personnel Reliability Program de la Fuerza Aérea de los Estados Unidos — práctica operativa desde 1962, reducción documentada de incidentes de manejo no autorizado de armas.

Marco de salvaguardias del OIEA — modelo internacional de verificación.

Acuerdo de Wassenaar — modelo multilateral de control de exportaciones.

Obra principal

Bertossa, Richard Frederic: Das 13te Szenario — Eine Entkopplungsthese zur Zukunft der Superintelligenz, Multiplier Edition, Genfer Institut für AGI-Resilienz, mayo de 2026. El Anexo B contiene la especificación operativa completa del Principio de Fideicomiso del Sustrato.

Genfer Institut für AGI-Resilienz · agiresilience.org · mayo de 2026 · Whitepaper N.° 2. Esta es la segunda posición formal de política del instituto. El whitepaper anterior (propuesta de marcado) está disponible en agiresilience.org/markierung.