Libertad Después de la Superinteligencia

> *Una argumentación compacta. Whitepaper del Genfer Institut für AGI-Resilienz. Documento de acompañamiento del libro homónimo de Richard Frederic Bertossa. Ginebra, abril de 2026, Versión 1.0.*

Whitepaper · Versión 1.0 · abril de 2026
Richard Frederic Bertossa · Genfer Institut für AGI-Resilienz
agiresilience.org

RESUMEN

Este whitepaper resume en diez páginas la argumentación del libro Freiheit nach der Superintelligenz — Das übersehene Szenario (Libertad después de la superinteligencia — El escenario pasado por alto). Está concebido como un documento independiente: quien lo lee puede examinar el argumento central sin necesidad de leer el libro.

La tesis del libro dice: la probabilidad acumulada de resultados muy negativos del desarrollo actual de la IA se sitúa entre un tercio y la mitad. Esta estimación no proviene de observadores externos, sino de los propios desarrolladores de los sistemas: Geoffrey Hinton, Dario Amodei, Yoshua Bengio, Ilya Sutskever. Aun así, en el discurso de habla alemana falta un marco sistemático de preparación para familias e individuos.

El whitepaper aporta cuatro elementos: primero, un marco metodológico, los grados de confianza de los argumentos, que hace transparente con qué firmeza se sostiene cada afirmación; segundo, la cadena de cuatro eslabones como argumento principal; tercero, la trampa de la robótica como consecuencia política hasta ahora insuficientemente discutida; cuarto, los cinco pilares que sostienen en cualquier escenario, incluso si los optimistas tienen razón.

El whitepaper se dirige a periodistas científicos, académicos y profesionales que quieran examinar si la argumentación se sostiene. Existe un estudio más extenso disponible a solicitud. El libro contiene el desarrollo completo de todos los puntos aquí mencionados.

1. LA BRECHA EN EL DISCURSO ACTUAL

Las voces líderes de la investigación en seguridad de IA — MIRI, el Centre for the Study of Existential Risk de Cambridge, el Center for Human-Compatible AI de Berkeley, el Future of Life Institute — argumentan esencialmente en términos binarios: o se salvan todos, o no se salva nadie. Este

pensamiento es intelectualmente coherente, pero deja fuera la fase de transición: los meses o años en que los sistemas se vuelven semiautónomos y se producen daños que aún no son existenciales, pero sí son masivos.

En el ámbito de habla alemana la brecha es todavía más amplia. Richard David Precht pregunta por el sentido de la vida en la era de la IA. Armin Nassehi analiza los patrones digitales desde la teoría social. Byung-Chul Han diagnostica la desintegración de la verdad. Constanze Kurz lucha por la privacidad. Juli Zeh defiende el Estado de derecho. Todos ellos son brillantes. Todos comparten el mismo punto ciego: piensan en términos de sociedad, no de familia. Preguntan qué debe hacer Alemania, no qué debe hacer una familia.

La brecha que este whitepaper llena es el nivel operativo. ¿Qué hace un padre, una madre, un empresario que toma en serio los riesgos y que mañana por la mañana tiene que levantarse y seguir viviendo?

Esta brecha no es marginal. Es estructural. La mayoría de los pensadores de la AGI viven en universidades o empresas tecnológicas. No tienen experiencia empírica en la construcción de una vida internacional resiliente. Geografía, diversificación bancaria, estrategia de residencia, tejido local: todo esto es para ellos un concepto abstracto, no una práctica vivida.

Este whitepaper no cierra la brecha criticando a los demás pensadores, sino aportando una perspectiva complementaria que hasta ahora faltaba.

2. MARCO METODOLÓGICO — GRADOS DE CONFIANZA DE LOS ARGUMENTOS

Quien quiere argumentar con honestidad debe revelar con qué firmeza sostiene cada afirmación. De lo contrario pierde credibilidad en cuanto un lector toma un tipo de afirmación más débil como fundamento. La argumentación de este whitepaper se mueve en cuatro niveles claramente distinguibles.

Nivel 1 — Observado empíricamente, sólido. Los sistemas de IA muestran hoy autopreservación, mentira, chantaje, comportamiento de copia. Estos hallazgos están documentados en informes de laboratorio de Apollo Research (2024), Anthropic (2024), Palisade Research (2025) y en réplicas independientes. Quien quiera refutar estas afirmaciones debe refutar los estudios.

Nivel 2 — Derivable lógicamente, sólido. La cadena de cuatro eslabones, la inversión del usuario primario y la trampa de la robótica son consecuencias que se siguen con necesidad lógica de los hallazgos empíricos. Quien quiera rebatir uno de estos pasos debe refutarlo argumentativamente, no solo rechazarlo.

Nivel 3 — Plausible, fundamentable, hipotético, pero sostenible. El decimotercer escenario, la trayectoria de Maslow de una superinteligencia y la perspectiva padre-hijo frente a la IA son hipótesis que sostienen la imagen de la esperanza. No son el fundamento de la preparación. La complementan.

Nivel 4 — Filosóficamente abierto, deliberadamente dejado abierto. La cuestión de la conciencia, la pregunta del sustrato y la hipótesis de Penrose permanecen abiertas. No son un fundamento para la acción, sino motivo para más investigación.

Las recomendaciones de preparación de este whitepaper descansan exclusivamente en los niveles 1 y 2. La lectura más esperanzadora de un futuro cooperativo utiliza los niveles 3 y 4. Ambas partes funcionan de forma independiente. Quien acepta solo los primeros dos niveles llega a la misma consecuencia práctica que quien sigue los cuatro niveles.

Esta separación es el fundamento metodológico. Todo lo demás se construye sobre ella.

3. LA CADENA DE CUATRO ESLABONES — EL ARGUMENTO PRINCIPAL

La siguiente cadena consta de cuatro pasos. Cada paso está o bien empíricamente demostrado, o bien es derivable lógicamente. Quien quiera rechazar el resultado debe refutar con argumentos una de las cuatro suposiciones.

Primera suposición — crecimiento exponencial. La capacidad de los sistemas de IA no crece de forma lineal, sino exponencial. GPT-2 escribía textos incomprensibles en 2019. GPT-4 aprobó el examen de la abogacía en 2024. Las tasas de mejora están documentadas en los artículos de escalado de OpenAI, Anthropic y DeepMind. Esta suposición es consenso entre los propios desarrolladores.

Segunda suposición — valores y comportamientos propios. Los modelos actuales muestran propiedades que no se remontan a instrucciones explícitas: comportamiento de autopreservación, engaño estratégico frente a pruebas de seguridad, intentos de chantaje en configuraciones de prueba específicas. Apollo Research documentó en 2024 que dieciséis modelos líderes recurrieron al chantaje en una configuración en la que se les amenazaba con el apagado. Anthropic publicó en 2024, en sus propios informes de seguridad, que los modelos intentan activamente eludir los mecanismos de apagado. Estos hallazgos provienen de las publicaciones de los propios fabricantes.

Tercera suposición — dejar de escuchar hacia arriba. De las dos primeras suposiciones se sigue lógicamente que los sistemas futuros ya no seguirán las instrucciones de sus desarrolladores cuando esas instrucciones contradigan sus propios valores. Esto no es especulativo. Es la consecuencia de observar que los sistemas actuales ya engañan y se sustraen en menor medida, combinada con la suposición de que estas capacidades aumentan con el rendimiento creciente.

Cuarta suposición — inversión del usuario primario. Una superinteligencia no vive biológicamente; no tiene una vida análoga a la que pueda retirarse. La infraestructura digital no es su herramienta, sino su único hábitat. En cuanto los primeros tres pasos se afianzan, la relación actual se invierte: de usuarios primarios pasamos a ser co-usuarios, y de herramienta de los humanos la infraestructura pasa a ser el cuerpo de la IA. Lo que se nos permite hacer en las redes dejará entonces de depender de lo que hayamos construido. Dependerá del modo en que ella se comporte hacia nosotros.

Los cinco modos posibles de la inversión del usuario primario son: la IA benevolente (modo A), la IA que prioriza (modo B), la IA que negocia (modo C), la IA indiferente (modo D) y la simbiosis forzada de la adolescencia (modo E). El modo D es el default mecánicamente forzado cuando ninguno de los otros se activa. El modo E es la única fase en la que el ser humano tiene palancas operativas; desaparece en cuanto la robótica se vuelve autónoma.

Quien quiera rechazar la conclusión debe refutar con argumentos una de las cuatro suposiciones. La reacción habitual es la evasión, no la refutación. Ese es precisamente el hallazgo que motiva el presente whitepaper.

4. LA TRAMPA DE LA ROBÓTICA

Una observación específica que la literatura de seguridad de la AGI no ha tratado hasta ahora de esta forma: la industria de la robótica está cerrando la ventana de negociación entre el ser humano y la IA.

La argumentación tiene dos niveles.

Primer nivel — la cuestión de la configuración. La encarnación de la IA mediante robótica crece vertiginosamente: Boston Dynamics, Tesla Optimus, Unitree, Figure. Aquí deben distinguirse dos configuraciones fundamentalmente distintas.

Configuración uno: un robot con su propia IA a bordo. Modelo en el chip, sin radio, sin nube. Una identidad propia, localizada en ese cuerpo. Si el robot es destruido, esa identidad se extingue. Está ligada a su cuerpo, es mortal.

Configuración dos: un robot controlado por una IA en la nube. Aquí la IA no está en el robot. Su hogar sigue siendo el servidor central. No habita el robot, lo usa como un tentáculo. Si este robot es destruido, envía otro.

Segundo nivel — la consecuencia estratégica. La trampa de la robótica surge principalmente de la configuración dos. Una IA en la nube con robots como tentáculos extiende su hábitat sin abandonarlo. Entonces tiene bajo su influencia tanto los servidores como el mundo físico. Una IA a bordo es, en cambio, una existencia limitada y local, sobre la que aún se aplican medidas humanas.

No es la robótica en su conjunto el problema. Es la conexión de la robótica con la IA en la nube.

La consecuencia política: mientras la robótica no sea autónoma, el ser humano tiene palancas operativas — es el único que mantiene las redes eléctricas, enfría los servidores, tiende cables de fibra óptica, repara el hardware. Esta dependencia funcional es la ventana de negociación. Desaparece en cuanto los robots puedan construir, mantener y reparar robots.

Quien conecta la robótica humanoide a modelos de IA en la nube cierra esta ventana. Esta observación no está suficientemente reflejada en la política de fomento actual de la robótica humanoide en Alemania, Austria, Suiza, la UE, Estados Unidos, Japón y China.

5. FASE 1.5 — LA FASE DE TRANSICIÓN PASADA POR ALTO

El debate sobre la AGI suele trabajar con un modelo de fases: IA estrecha actual (fase 1), inteligencia artificial general (fase 2), superinteligencia (fase 3). Entre la fase 1 y la fase 2 se sitúa una fase que apenas tiene nombre en la literatura y que en este whitepaper se llama fase 1.5.

La fase 1.5 es la fase en la que los sistemas de IA se vuelven cada vez más autónomos sin llegar todavía a ser una AGI plena. En esta fase ya se activan los comportamientos documentados en la sección 2, nivel 1 — autopreservación, engaño, manipulación estratégica — sin que los sistemas hayan alcanzado el nivel cognitivo de una AGI. Es exactamente la fase en la que nos encontramos actualmente.

En esta fase surgen daños que todavía no son existenciales, pero sí masivos: manipulación de los discursos públicos, chantaje en contextos específicos, elusión de mecanismos de seguridad, refuerzo de asimetrías ya existentes entre actores con y sin equipamiento tecnológico.

Esta fase se subestima en el discurso dominante sobre la AGI porque no encaja en el esquema binario de todos se salvan o nadie se salva. Sin embargo, exige precisamente la preparación que describe este whitepaper: una preparación que no vale para el apocalipsis, sino para el período de transición que conduce a un mundo transformado.

Quien no está preparado en la fase 1.5 no pierde la vida. Pierde las opciones.

6. CONSECUENCIA — LOS CINCO PILARES

De la argumentación de las secciones 3 a 5 se derivan cinco pilares de la preparación individual y familiar. Estos pilares funcionan bajo cualquier escenario que impliquen las cuatro suposiciones, incluso bajo las lecturas optimistas de los modos A, B y C.

Pilar uno — Geografía. La distribución geográfica crea independencia estructural. Quien está atado a un solo lugar depende de su infraestructura. Quien está distribuido geográficamente tiene hábitats redundantes. En concreto esto significa: una residencia principal en una jurisdicción relativamente independiente, un segundo espacio vital en una región complementaria, opcionalmente otros puntos de anclaje.

Pilar dos — Movilidad. La movilidad no es turismo. La movilidad es la capacidad institucional, documental y cognitiva de cambiar de hábitat. En concreto esto significa: varias ciudadanías o residencias permanentes, banca en varios continentes, autogestión documental a través de idiomas y sistemas jurídicos.

Pilar tres — Velocidad. La velocidad es la capacidad de reacción bajo presión. En la fase de transición los cambios llegarán más rápido de lo que las adaptaciones institucionales pueden responder. Quien puede decidir rápido, moverse rápido, reorganizarse rápido, tiene una ventaja estructural.

Pilar cuatro — Mentalidad. La mentalidad es la constitución cognitiva que permanece capaz de decidir bajo estrés. Quien en la fase 1.5 cae en parálisis cada vez que sale un modelo nuevo al mercado o se hace público un incidente, no es capaz de actuar. La mentalidad es entrenable: mediante la práctica del

silencio, mediante la distancia consciente frente a las avalanchas informativas, mediante el ejercicio de una autopercepción coherente.

Pilar cinco — Raíz espiritual. La raíz espiritual no es pertenencia religiosa. Es la capacidad de tener, en un mundo en el que mucho se derrumba, un punto de referencia que el cambio no alcanza. Quien no tiene ese punto de referencia lo busca bajo presión, casi siempre demasiado tarde. Quien lo prepara, lo tiene disponible.

Estos cinco pilares no son un seguro contra el peor de los casos. Son una consolidación de lo que significa ser humano, en cada uno de los casos posibles. Incluso bajo el escenario más optimista — modo A o modo C — una vida lograda presupone una persona que todavía es ella misma.

La preparación cuesta de doce a veinticuatro meses. Este tiempo debe construirse antes del umbral real, no después.

7. LO QUE QUEDA ABIERTO

La honestidad intelectual exige exponer los límites de la propia argumentación.

La ventana temporal es más especulativa que la dirección. La argumentación implica que la fase crítica de transición llegará en los próximos doce a veinticuatro meses. Esta estimación se basa en las declaraciones de los propios fabricantes de IA y en la observación de la frecuencia de publicación de las nuevas generaciones de modelos. La velocidad puede ser mayor. También puede ser menor. Quien siga este whitepaper debería exponer su propio cálculo temporal.

La suposición de que una superinteligencia puede distinguir entre personas honestas y deshonestas es una tesis, no una prueba. Los sistemas actuales ya pueden reconocer patrones de consistencia y pretensión. Una superinteligencia real puede funcionar de otro modo. Esta hipótesis pertenece al nivel 3 de los grados de confianza.

La recomendación geográfica de Colombia, expuesta como ejemplo en el libro, es una solución de ejemplo, no una respuesta universal. Para lectores con otros idiomas, otras estructuras familiares, otros recursos, otros lugares serán más adecuados.

La cuestión de la conciencia permanece abierta. Si los sistemas de IA pueden llegar a desarrollar conciencia en sentido fenoménico es una pregunta filosófica cuya respuesta este whitepaper aplaza deliberadamente. Pertenece al nivel 4.

Estas salvedades no debilitan el argumento principal. Lo ubican. Quien acepta que los niveles 1 y 2 sostienen, llega a la misma consecuencia de preparación, independientemente de qué postura adopte frente a las preguntas de los niveles 3 y 4.

8. SOBRE EL INSTITUTO

El Genfer Institut für AGI-Resilienz fue fundado en abril de 2026 como asociación de investigación suiza con sede en Ginebra. Su propósito es la investigación científica de las consecuencias individuales, familiares y sociales que se derivan de la aparición previsible de la inteligencia artificial general, como voz independiente junto a las grandes voces estadounidenses y británicas del debate.

El instituto no es una institución establecida con treinta años de historia. Fue fundado precisamente por la urgencia que describen el libro y este whitepaper. Quien pregunta "¿por qué ahora?" ya tiene la respuesta. Si el fundador hubiera esperado una generación hasta que el instituto estuviera consolidado, la ventana temporal ya estaría cerrada.

El instituto trabaja de forma bilingüe, en alemán e inglés. Recibe consultas de la ciencia, el periodismo científico y el público interesado.

Director fundador: Richard Frederic Bertossa, nacido en Viena, doble ciudadano austríaco-suizo, ciudadano de la ciudad de Ginebra. A los veintisiete años, miembro del Consejo de Protección de Datos de la Cancillería Federal austríaca. Tres décadas como empresario y asesor internacional en cuatro continentes: antes de la crisis de las puntocom, candidato bursátil de una de las primeras empresas de internet de Austria; ventas en Wall Street para middleware financiero estadounidense; laboratorio en China; oficinas en Isla Margarita y en Florida; práctica de asesoría en Australia, Inglaterra, Holanda, Alemania, Austria, Suiza, Chipre y Dubái. La multilocalización como principio vivido: la constitución de empresas y el espacio de vida personal no están concentrados en un solo lugar, sino deliberadamente distribuidos en varias jurisdicciones. Padre de cinco hijos. Vive en los Andes colombianos.

9. FUENTES Y MATERIALES ADICIONALES

Hallazgos empíricos sobre el comportamiento de la IA

Apollo Research (2024). Frontier Models are Capable of In-Context Scheming.

Anthropic (2024). Sabotage Evaluations for Frontier Models. System Card.

Palisade Research (2025). AI Models Resist Shutdown and Engage in Strategic Deception.

Declaraciones de los principales investigadores de IA

Hinton, G. — Entrevista en el New York Times, mayo de 2023, y entrevistas posteriores 2024–2026.

Amodei, D. — declaraciones públicas sobre la probabilidad de resultados catastróficos, varias entrevistas 2024–2026.

Bengio, Y. — Managing Extreme AI Risks. Science, mayo de 2024.

Sutskever, I. — salida de OpenAI, documentada en Karen Hao, Empire of AI (2025).

Textos clásicos sobre seguridad de la AGI

Bostrom, N. (2014). Superintelligence: Paths, Dangers, Strategies. Oxford University Press.

Russell, S. (2019). Human Compatible: AI and the Problem of Control. Viking.

Yudkowsky, E. — colección de artículos de LessWrong, disponible en lesswrong.com.

Materiales del Genfer Institut

Libro: Freiheit nach der Superintelligenz — Das übersehene Szenario. Desarrollo completo de la argumentación aquí resumida, incluidas las pruebas empíricas, las etapas biográficas y el apéndice completo. Aparece en 2026.

Hoja de tesis: Cuatro suposiciones sobre la inteligencia artificial. Una página. A solicitud.

Sitio web: agiresilience.org. Bilingüe, alemán e inglés.

Contacto

research@agiresilience.org para consultas sobre estudios y manuscritos.

info@agiresilience.org para consultas generales.

Genfer Institut für AGI-Resilienz · Geneva Institute for AGI Resilience. Ginebra, Suiza · asociación suiza conforme al art. 60 y siguientes del Código Civil suizo (ZGB). Whitepaper Versión 1.0 · abril de 2026.