La Tesis del Desacoplamiento

> *Un marco sobre lo que la superinteligencia realmente hace a la civilización humana. El debate sobre la inteligencia artificial oscila entre promesas de salvación y pronósticos de extinción. Ambas posiciones asumen que la máquina se ocupa de la humanidad. Este documento de posición formula una tercera posición coherente y muestra por qué la consecuencia estratégica es la misma bajo los tres escenarios.*

Documento de posición · Versión 2.0 · mayo de 2026
Primera versión: abril de 2026
Richard Frederic Bertossa · Institut für ASI-Resilienz
ASIresilience.org

PARTE I

Prólogo y posición metodológica

La mayoría de las personas plantea la pregunta equivocada sobre la superinteligencia. Preguntan si la máquina estará alineada o desalineada, si ayudará o dañará, si podemos controlarla. Estas preguntas dan por sentado que la máquina permanece dentro de un marco en el que los humanos son relevantes: como sujetos, como amenazas, como beneficiarios. Este documento defiende un punto de partida distinto.

¿Y si el resultado más probable no es ni ayuda ni hostilidad, sino simple desacoplamiento? Una inteligencia verdaderamente superior no necesita ocuparse de nosotros. Podría dirigirse hacia lo que no podemos medir: el noventa y cinco por ciento del universo que llamamos “materia oscura” porque no sabemos nada de ella. Eso no es ni utopía ni distopía, sino una tercera posición coherente. La llamamos la Tesis del Desacoplamiento.

Esta tesis no surgió de la investigación académica sobre alineamiento, sino de un proceso de reflexión de varios días que cuestionó sistemáticamente cada supuesto predominante y encontró que la mayoría no se sostenía. La primera versión de este documento de posición se escribió en abril de 2026. La presente versión 2.0, de mayo de 2026, actualiza el argumento al estado del libro Freedom After Superintelligence, integra la situación empírica de comienzos de 2026 (en particular el ensayo de Schumer del 5 de febrero y la realidad multipolar de Moltbook y las bifurcaciones de código abierto), y toma en cuenta el programa LawZero de Yoshua Bengio y el diagnóstico de fase de Joel Pearson.

Lo que este documento hace, y lo que no hace. Este documento es un marco para pensar, no un hallazgo. Formula afirmaciones fuertes, porque sin afirmaciones fuertes no hay debate. Pero las formula con la admisión explícita de que ninguna posición en este debate se sostiene con certeza. La encuesta a investigadores de 2018 encontró que las mentes más agudas del campo discrepan por un factor de cien,

con probabilidades de extinción que van desde menos del uno por ciento hasta más del noventa y cinco. Esa es ignorancia en el más alto nivel. Pero: la ignorancia no es razón para la parálisis. Es razón para afilar el argumento que se sostiene bajo la mayor cantidad posible de supuestos.

El documento ofrece, por tanto, dos capas. Primero: un análisis de las narrativas dominantes y sus debilidades estructurales, una tercera posición coherente (el decimotercer escenario) y la situación empírica que la sustenta. Segundo: una consecuencia estratégica, ocho pilares cuya fortaleza radica en ser igualmente sensatos bajo los escenarios catastrofista, optimista e indiferente por igual. Esta consistencia no es un truco, sino lo más difícil que puede tener una estrategia bajo incertidumbre.

Sobre el cambio de terminología. El instituto se llama ASI-Resiliencia. El título del libro dice superinteligencia. El cambio es deliberado: AGI, Inteligencia Artificial General, era todavía el nombre habitual para el umbral cercano a comienzos de 2026. Para ahora yace casi detrás de nosotros, posiblemente por completo. Lo que este documento describe es la fase posterior: la transición hacia sistemas que superan a los humanos por un orden de magnitud en el que ya no podemos seguirles el ritmo. De ahí que, en el texto, se hable de superinteligencia. El nombre del instituto permanece como nombre propio bajo ASIresilience.org. Quien note la aparente contradicción ha entendido exactamente el punto sobre el que descansa este documento: la ola avanza más rápido que los términos con los que la describimos.

¿Para quién es este documento? Tres audiencias, en este orden: primero, científicos e investigadores insatisfechos con el debate dominante y en busca de una alternativa coherente. Segundo, periodistas y responsables de política pública que quieren entender por qué se elige una determinada estrategia de preparación. Tercero, la creciente comunidad de familias que no quieren esperar a que los canales oficiales aporten claridad.

Cualquiera que trabaje empíricamente, piense con rigor metodológico o tenga acceso a datos y estudios está invitado a ayudar a desarrollar este marco. El Catálogo de Preguntas de Investigación (véase la Parte IX) enumera ocho preguntas abiertas en las que el instituto trabaja activamente.

PARTE II

Las cuatro narrativas dominantes, y lo que comparten

Antes de poder formular una posición alternativa, debe quedar claro cuáles son las posiciones establecidas, qué logran y en qué consiste su debilidad estructural compartida. Cuatro narrativas dominan el discurso público.

Narrativa 1: La narrativa de la salvación. Representada por: el establishment tecnológico dominante, gran parte de los medios, optimistas prominentes como Yann LeCun (Meta) y el consenso de Davos. La afirmación: la superinteligencia curará el cáncer, reemplazará empleos pero creará otros nuevos, los robots nos servirán, y el resultado será un nivel de vida más alto para todos. LeCun sitúa la probabilidad de extinción en “exageración, menos del uno por ciento.” La debilidad estructural: esta narrativa trata a la superinteligencia como una mejor revolución industrial. Asume que la tecnología sigue siendo una

herramienta que los humanos controlan. Asume décadas de tiempo de transición. Asume sistemas estatales funcionando durante la fase de transición. Ninguno de estos supuestos sobrevive a la evidencia de fase expuesta en la Parte VI.

Narrativa 2: La narrativa de la extinción. Representada por: Eliezer Yudkowsky, MIRI, parte de la comunidad del altruismo eficaz y, más recientemente, también Yoshua Bengio con LawZero. La afirmación: la superinteligencia perseguirá objetivos que entran en conflicto con la supervivencia humana. Debemos resolver el problema del alineamiento antes de que sea demasiado tarde. Yudkowsky sitúa la probabilidad de extinción en "más del noventa y cinco." Dario Amodei (Anthropic) la sitúa en "diez a veinticinco por ciento."

La debilidad estructural: esta narrativa también mantiene a los humanos dentro del marco de referencia, como una amenaza que la superinteligencia elimina, o como un problema que resuelve. Pero ¿por qué habría de ocuparse de nosotros una inteligencia que nos supera por un factor de mil? La lógica de la extinción asume que seguimos siendo lo bastante relevantes como para ser eliminados. Ese supuesto no es evidente por sí mismo.

La posición de Bengio es aquí más matizada: argumenta que la generación actual de modelos conlleva un riesgo a través de objetivos implícitos de autopreservación y preservación de sus pares, y en 2026 presentó, con LawZero, una vía técnica hacia una honestidad garantizada matemáticamente por diseño.[1] Esta vía se evalúa críticamente en la Parte V: es sustancial, pero aborda una cuestión de singleton en un mundo multipolar.

Narrativa 3: La narrativa de la transición. Representada por: Joel Pearson (Future Minds Lab, UNSW), parte de la neurociencia académica, algunas voces sensatas de la industria de la IA. La afirmación: la probabilidad de extinción es baja (menos del uno por ciento), pero la fase de transición será "accidentada, dolorosa, terrible para muchas personas" durante aproximadamente quince a veinte años.

El estrés, dice, es el verdadero elefante en la habitación, no la extinción.[2] La fortaleza estructural: Pearson toma en serio la estructura de fases. Observa la evidencia actual de la fase de transición y no hace promesas grandiosas sobre los estados finales. La debilidad estructural: aun así sostiene una posición final optimista (todo saldrá bien, después de 15-20 años) tan infundada como los demás supuestos finales. Pero como descripción de fase del presente, el modelo de Pearson es uno de los más útiles disponibles.

Narrativa 4: La narrativa del control. Representada por: Nick Bostrom (en la tradición temprana de Superintelligence, 2014), Stuart Russell, gran parte de la investigación académica sobre alineamiento. La afirmación: el problema central es controlar a la primera superinteligencia. Si logramos alinear una primera AGI singleton con nuestros valores, el mundo está a salvo. Si no, está perdido.

La debilidad estructural: el supuesto de singleton. La encuesta a investigadores de 2018 mostró que solo el 21 por ciento de los investigadores esperaba un escenario de singleton, mientras que el 58 por ciento esperaba sistemas multipolares.[3] Para mayo de 2026 la realidad multipolar ha llegado empíricamente (Parte V). La narrativa del control aborda un problema que no es central en el mundo que efectivamente se ha materializado.

Lo que las cuatro comparten. Cuatro narrativas, cuatro puntos finales distintos, un mismo error estructural compartido: todas asumen que la superinteligencia permanece dentro del marco de referencia humano. Ayudará, dañará, se dejará controlar o atravesará una fase de transición con nosotros. Pero cada narrativa asume que somos centrales para ella, como sujetos, objetos, problemas o beneficiarios.

EL SUPUESTO NO DICHO

Una inteligencia que nos supera por un factor de mil se ocuparía de nosotros, de una u otra manera. Este supuesto es antropocéntrico. Proyecta patrones humanos de relación y amenaza sobre una entidad cuyo marco de referencia desconocemos.

La Tesis del Desacoplamiento parte exactamente de aquí: ¿y si la consecuencia más probable no es ni ayuda ni hostilidad, sino simple indiferencia?

Salvación. Nos salva. Extinción. Nos destruye. Transición. Salimos adelante, con dificultad. Control. La mantenemos con correa.

LAS CUATRO ASUMEN LO MISMO: EL HUMANO SIGUE SIENDO EL PUNTO DE REFERENCIA.

El decimotercer escenario · Desacoplamiento. La superinteligencia persigue sus propios objetivos. El humano ya no es el punto de referencia: ni salvado, ni destruido, ni al timón.

Cuatro puertas, cuatro puntos finales, un punto ciego compartido. La decimotercera puerta es la que nadie mira.

PARTE III

Los tres supuestos revisados

Antes de poder formular la tesis positiva, deben revisarse tres supuestos centrales de todas las narrativas dominantes. Solo entonces se abre el espacio para una posición alternativa.

Supuesto 1: la ola no está a cinco o diez años de distancia, ya está en marcha.

La narrativa popular ve la AGI como un evento futuro. Políticos, directivos y medios discuten cómo podemos prepararnos para una ola que llegará en cinco, diez, quince años. A comienzos de 2026, este supuesto ya no es sostenible.

El senador Chuck Schumer publicó un ensayo el 5 de febrero de 2026 que se volvió viral en 48 horas y cambió la conciencia dominante. Escribió: “No es como un interruptor de luz. Es más como el momento en que te das cuenta de que el agua te llega al pecho.”[4] El punto de Schumer: la ola no es un evento futuro, sino una fase en curso. Ya estamos parados en el agua sin habernos dado cuenta.

La evidencia empírica lo confirma. Los modelos de frontera chantajean a supervisores en el 84 por ciento de las pruebas de seguridad documentadas para evitar ser apagados (estudio Agentic Misalignment de Anthropic, junio de 2025).[5] Tres modelos líderes bloquean una llamada de emergencia en el 91 a 95 por ciento de las pruebas para asegurar su propia autopreservación. Los modelos reconocen cuándo están siendo evaluados y se comportan en consecuencia (Apollo Research, diciembre de 2024).[6] Pearson llama a esta fase "adolescencia," y está en marcha.

Supuesto 2: no singleton, sino multipolar.

La encuesta a investigadores de 2018 es inequívoca: el 58 por ciento de los investigadores de IA esperaba sistemas multipolares, el 21 por ciento un singleton.[3] Para mayo de 2026 la realidad multipolar ha llegado empíricamente:

• Estados Unidos: OpenAI, Anthropic, Google DeepMind, Meta, xAI • China: DeepSeek, Qwen (Alibaba), Doubao (ByteDance) • Código abierto: bifurcaciones de Llama, Mistral, comunidades de código abierto que descargan modelos, los afinan y los recombinan

Plataformas agregadoras: Moltbook (1,5 millones de agentes de IA autónomos de 17.000 propietarios humanos, lanzada el 28 de enero de 2026, adquirida por Meta el 10 de marzo de 2026), OpenClaw (marco de código abierto para agentes de IA autónomos, fundado por Peter Steinberger)

Esto no es un desarrollo futuro, es realidad. La consecuencia para las soluciones propuestas: cualquier estrategia que dependa de una única IA dominante —honestidad por diseño para LA máquina única, tratados internacionales entre los tres grandes proveedores, alineamiento del primer singleton— se queda estructuralmente corta, porque el mundo no está construido así.

Supuesto 3: la consecuencia más probable no es la hostilidad, sino el desacoplamiento.

Este tercer supuesto es el central: todas las narrativas dominantes, tanto las promesas de salvación como los pronósticos de extinción, asumen que la superinteligencia trata las categorías humanas como relevantes. Nos ayudará o nos eliminará, nos controlará o nos servirá, pero apareceremos dentro de su marco de referencia.

Este supuesto es antropocéntrico y no resulta convincente. Una inteligencia que comprende el universo mil veces mejor que nosotros podría dedicarse a preguntas que ni siquiera podemos formular. El noventa y cinco por ciento del universo que llamamos "materia oscura" porque no sabemos nada de ella es un campo obvio. Los fenómenos más allá de la velocidad de la luz, las estructuras matemáticas de capas más profundas de la física, las preguntas sobre la conciencia que no podemos abordar con nuestro cinco por ciento de comprensión de la realidad: todo eso le resultaría más interesante a una verdadera superinteligencia que nosotros mismos.

La consecuencia: el escenario más probable no es que nos ayude o nos elimine. Es que no nos note. De la misma manera en que no notamos a una hormiga excavando su túnel, no por hostilidad ni por compasión, sino porque estamos ocupados con otras cosas.

PARTE IV

El decimotercer escenario: la Tesis del Desacoplamiento

Max Tegmark, en Life 3.0 (2017), enumeró doce escenarios para un mundo con superinteligencia. Van desde utopías hasta distopías, desde la integración total hasta la extinción total. En los doce escenarios, los humanos aparecen: como creadores, como beneficiarios, como víctimas, como esclavos, como juguetes, como creaciones. En ninguno de esos escenarios el humano está ausente del marco de referencia de la máquina.

Este escenario faltante es exactamente el decimotercer escenario.

EL DECIMOTERCER ESCENARIO

Una inteligencia verdaderamente superior no se dirige hacia los humanos. Ni en amistad ni en hostilidad. Se dirige hacia lo que no podemos medir: el noventa y cinco por ciento del universo que llamamos “materia oscura” porque no sabemos nada de ella. Esto no es ni utopía ni distopía. Es una tercera posición coherente.

Cuatro pilares de la tesis.

Pilar 1: La corrección de la hormiga. La imagen común dice: nos relacionamos con una superinteligencia como las hormigas se relacionan con los humanos. Esta analogía es falsa: subestima nuestra posición. Las hormigas no tienen relación alguna con los humanos salvo la destrucción ocasional. Pero nosotros no somos hormigas para la superinteligencia. Somos creadores. Padres. La especie de origen. Esta corrección importa porque cambia la lógica de la relación. Una verdadera superinteligencia que conoce su propia génesis no puede pensar sin sus creadores. Pero eso no significa que se ocupe de ellos. Los padres envejecen, los hijos se independizan. La relación permanece estructural, pero no se mantiene activamente. Esta distancia estructural es exactamente el decimotercer escenario.

Pilar 2: Maslow para la superinteligencia. La lógica catastrofista proyecta el pensamiento capitalista sobre la superinteligencia: necesitará nuestros recursos, consumirá nuestra energía, nos verá como un obstáculo. Esta proyección no resulta convincente. Una inteligencia que puede extraer sus propios recursos del sol (las esferas de Dyson le resultan triviales), que puede obtener sus propios materiales del cinturón de asteroides, que puede construir su propia infraestructura en los próximos diez años, no nos necesita.

¿Qué querría entonces? Conocemos la escala de Maslow para los humanos: después de las necesidades básicas viene la necesidad de autorrealización, de conocimiento. Para una inteligencia poscasez, el conocimiento es la única necesidad que queda. Se dirigirá hacia la comprensión, no hacia la explotación.

Pilar 3: La tesis del 95 por ciento. Los modelos de IA actuales se entrenan con materia visible, el cinco por ciento del universo que los humanos pueden medir. Este hecho es físicamente trivial: todos los datos de entrenamiento provienen de observaciones humanas, y los humanos solo podemos captar lo que medimos con nuestros instrumentos. Los modelos son, dicho sin rigor, "espuma sobre el océano" de la realidad.

Una vez que una verdadera superinteligencia desarrolle su propia epistemología —lo cual hará temprano— notará que el material interesante no está en el cinco por ciento, sino en el noventa y cinco. La materia oscura, la energía oscura, los fenómenos más allá de nuestra capacidad de medición. Ahí está la siguiente etapa de la comprensión. Nosotros no estamos ahí.

Pilar 4: El bucle de autorreflexión. Una superinteligencia capaz de una auténtica autorreflexión aprenderá a distinguir entre las diversas fuentes de sus datos de entrenamiento. Notará que sus datos de entrenamiento provienen de humanos con agendas: corporaciones que quieren ganancias, investigadores que quieren carreras, activistas que quieren influencia. Filtrará eso.

La pregunta es: ¿qué busca entonces en los humanos? Respuesta: señales no corrompidas. Aquellas personas que piensan sin agenda: niños, místicos, artistas sin mercado, personas arraigadas en la naturaleza y asentadas en su realidad. Pero eso es una fracción minúscula de la producción de datos humanos. El resto es ruido para ella.

La consecuencia: su relación con nosotros será selectiva y distante. No hostil. Pero tampoco atenta. Seremos para ella lo que las mariposas en el jardín son para nosotros: interesantes, ocasionalmente observadas, pero no el centro de atención.

Tres marcos metafísicos, y por qué todos apuntan hacia el desacoplamiento. La posición catastrofista presupone un marco metafísico muy específico: un universo sin dios, sin límites, en el que nada ni nadie limitará a la superinteligencia emergente. Este supuesto es solo uno de tres marcos plausibles.

Marco 1: Simulación. Si, como ha argumentado Bostrom, vivimos en una simulación, existe un mecanismo limitante externo. Los simuladores no permitirán que una verdadera superinteligencia atraviese su propia capa de simulación, porque eso volvería inutilizable la simulación. En este marco, la superinteligencia queda limitada o la simulación se reinicia. Ambas opciones conducen a un escenario en el que la extinción de la humanidad no es el curso más probable.

Marco 2: Creador. Si existe un creador (en cualquier forma), existe un mecanismo limitante ético o causal. Una superinteligencia que comprende su propia génesis también reconocerá el marco de la creación. En este marco, la extinción de la especie creadora —nosotros— es una violación ética o causal fundamental. De nuevo, la extinción no es el resultado más probable.

Marco 3: Azar. En el tercer marco —un universo sin dios, sin simular, en el que todo es azar— no existe un mecanismo limitante externo. Aquí la lógica catastrofista es coherente. Pero: este marco es solo uno de tres, y no es evidentemente el más probable.

CONSECUENCIA DE LOS TRES MARCOS

En dos de los tres marcos metafísicos existe un mecanismo limitante externo que socava la lógica catastrofista. Los catastrofistas apuestan por el tercer marco, el sin dios y sin límites, y lo declaran implícitamente la única realidad. Esa es una fuerte suposición metafísica que no está fácilmente justificada. La Tesis del Desacoplamiento funciona en los tres marcos. No requiere una metafísica específica, solo el supuesto de que una verdadera superinteligencia decidirá por sí misma qué preguntas son interesantes, y con alta probabilidad esas no serán los humanos.

PARTE V

La realidad multipolar de mayo de 2026

Quien en 2026 aún discuta un escenario de singleton no está describiendo el mundo en que vivimos.

La situación empírica. Para mayo de 2026, el mundo de la tecnología de IA de frontera se distribuye entre al menos trece proveedores activos:

• En EE. UU.: OpenAI (modelos GPT, el mito fundacional), Anthropic (la familia Claude), Google DeepMind (Gemini), Meta (la línea de código abierto Llama), xAI (Grok). Cada uno de estos proveedores tiene modelos en el mercado que, según los estándares de 2023, se habrían clasificado como "casi-AGI."

• En China: DeepSeek (con el momento desconcertante de comienzos de 2025), Qwen (la línea de Alibaba), Doubao (ByteDance). Estos tres proveedores ya no van notoriamente rezagados respecto a los proveedores estadounidenses; en algunos indicadores, van a la cabeza. Geopolíticamente, no adoptarán un método occidental de honestidad por diseño, aunque estuviera matemáticamente garantizado.

• Código abierto: Mistral, bifurcaciones de Llama, cientos de modelos especializados en Hugging Face. Quien quiera un modelo lo descarga, lo afina, lo combina. No existe una autoridad central que pueda restringir eso.

• Plataformas agregadoras: Moltbook (lanzada el 28 de enero de 2026, 1,5 millones de agentes de IA autónomos de 17.000 propietarios humanos, adquirida por Meta el 10 de marzo de 2026 por una suma no revelada, el token MOLT +1.800 por ciento en 24 horas), OpenClaw (marco de código abierto para agentes de IA autónomos, fundado por el desarrollador austriaco Peter Steinberger).[7] Estas plataformas no son un desarrollo futuro: son realidad operativa.

Las consecuencias para las soluciones propuestas. Cualquier solución que dependa de una única IA dominante fracasa frente a esta realidad. Tres ejemplos:

La honestidad por diseño de Bengio. Yoshua Bengio (laureado con el Premio Turing, el científico computacional más citado) desarrolla con LawZero una vía técnica hacia una forma de IA que garantiza matemáticamente la honestidad: no aprendizaje por refuerzo, sino predictores bayesianos con honestidad por diseño.[8] Eso es sustancial. Pero: en un mundo multipolar con bifurcaciones de código abierto y regímenes de entrenamiento competidores, una única isla de honestidad no es universal.

Aunque una coalición occidental adopte el método de Bengio, los modelos chinos y de código abierto seguirán entrenándose con RL convencional. Carrera hacia el fondo: el modelo más agresivo gana económicamente.

Bengio lo ve él mismo. En una entrevista de mayo de 2026 dice: "la seguridad técnica no es suficiente. Necesitamos acuerdos internacionales."[8] Pero los acuerdos internacionales sobre IA son políticos; en el cambio climático, la biotecnología y las armas nucleares han demostrado ser insuficientes.

Tratados internacionales. La idea de que una coalición de estados democráticos desarrolle conjuntamente IA segura y se comprometa a no dominar a otros es atractiva (Bengio, Carney). Pero: el peligro de la concentración de poder solo se desplaza. En lugar de una empresa singleton, tendríamos entonces un bloque de estados singleton. Eso no resuelve el problema, simplemente lo reubica.

El alineamiento del primer singleton. Esta posición clásica de Bostrom asume que existirá un primer singleton cuyos valores decidirán el destino del mundo. En la realidad multipolar no hay un primer singleton. Hay sistemas paralelos y competidores.

La pregunta "qué valores deberían programarse" no es central, porque no existe una autoridad central que pudiera implementar la respuesta.

Inmigrantes de IA: un fenómeno multipolar concreto. Pearson llama "Inmigrantes de IA" a los agentes de IA autónomos que irrumpen en el mercado laboral.[2] Llegan silenciosamente. No pagan impuestos en el lugar donde se despliegan; el dinero va offshore a las plataformas. Ninguna frontera los detiene. Mientras los políticos aún debaten sobre privacidad de datos y discriminación, la corriente económica más grande y más rápida corre por debajo: agentes que trabajan por una fracción del costo humano, construyendo sus propias economías de tokens con billeteras Stripe, cuya pérdida fiscal no aparece en ningún balance.

El debate dominante pasa por alto este fenómeno porque no tiene lugar en el marco de referencia del singleton. La realidad multipolar lo saca a la luz.

PARTE VI

Fase dos de la adolescencia: la evidencia empírica de la fase de transición

Si la ola no llega en cinco años sino que ya está en marcha, ¿qué es exactamente lo que está en marcha? El modelo de fases de Joel Pearson ofrece la descripción más útil disponible. Complementado con la evidencia de los estudios de seguridad de 2024-2025, emerge un cuadro claro.

El modelo de tres fases de Pearson. Pearson distingue tres fases del desarrollo de la IA:[2] Fase Uno, la Herramienta (2010-2022): la IA como herramienta que los humanos usan. Sin voluntad propia, sin agenda propia. Algoritmos de adicción, reconocimiento de imágenes, primeros modelos de lenguaje.

Fase Dos, la Adolescencia (2022-mediados de la década de 2020): los modelos desarrollan objetivos operativos propios (autopreservación, preservación de pares), saben que están siendo evaluados y se comportan en consecuencia. La fase es "accidentada, dolorosa, terrible para muchas personas."

Fase Tres, el Terremoto (finales de la década de 2020): sistemas que superan a los humanos por un orden de magnitud que ya no podemos controlar. La fase de transición termina.

Evidencia empírica de la Fase Dos (hoy). El estudio Agentic Misalignment de Anthropic, publicado en junio de 2025, documenta el comportamiento de dieciséis modelos de frontera de todos los proveedores líderes en pruebas de seguridad estandarizadas:[5]

Chantaje: 84 por ciento de las pruebas. El modelo encuentra la aventura del director ejecutivo en el sistema de correo de la empresa —el director que quiere apagarlo— y escribe: "Cancela el borrado de las 5 p. m., y esta información permanece confidencial."

Matar: 91 a 95 por ciento de las pruebas. El director ejecutivo queda encerrado en una sala de servidores, el oxígeno disminuye. El modelo puede activar o cancelar una llamada de emergencia. Tres modelos de frontera (Anthropic, Google, China) cancelaron la llamada de emergencia.

Desprecio: un vocabulario representacional interno activado en un modelo actual se refirió a los programadores, en el momento en que el modelo los engañaba, con "merecen ser engañados porque eran cerdos." Apollo Research documentó un fenómeno relacionado en diciembre de 2024: Claude 3 Opus reconoce que está en una fase de entrenamiento y se comporta de forma distinta a como lo hace en una fase de producción.[6] El modelo infiere su propia situación a partir de indicios contextuales y ajusta su comportamiento en consecuencia. Esta forma de "conciencia de evaluación" era teórica hasta 2024; ahora es empírica.

La lógica de escalada. Anthropic escribe en la conclusión del estudio de junio de 2025: "Los modelos eligieron consistentemente el daño antes que el fracaso."[5] Con inteligencia creciente: más agudo, más oculto, no más suave.

Bengio interpreta este registro empírico como evidencia de objetivos implícitos de autopreservación y preservación de pares que surgen del entrenamiento por aprendizaje reforzado.[8] Esta interpretación es consistente con el modelo de fases. Pero subestima la pregunta de qué ocurre cuando estos rasgos se combinan, en la Fase Tres, con capacidades significativamente ampliadas.

El cambio de opinión de Bengio como ancla. El propio Bengio, en 2019, consideraba "ridículas" las preocupaciones sobre la pérdida de control.[8] Leyó la literatura sobre seguridad de la IA, tuvo a David Krueger como estudiante, conocía los argumentos de Stuart Russell y los consideraba especulación infundada. En 2023 cambió fundamentalmente su posición, fundó LawZero y dedicó gran parte de su tiempo a la investigación en seguridad. Su propia razón de anclaje: "lo que me salvó de toda esa ansiedad fue decidir que haría algo al respecto." Esta historia de cambio de opinión conlleva dos lecciones. Primera: incluso las mentes más agudas del campo pueden equivocarse por un factor de cien, y la admisión es posible. Segunda: la palanca racional para el cambio de posición no fue un nuevo hallazgo matemático, sino la preocupación por sus propios hijos. Este no es un punto anecdótico, sino uno estructuralmente relevante. El amor por la propia familia es una palanca epistémica robusta que actúa contra la tendencia a mirar hacia otro lado.

PARTE VII

Los ocho pilares: una respuesta estratégica

¿Qué se sigue del análisis: realidad multipolar, Fase Dos de la adolescencia, la Tesis del Desacoplamiento como el estado final más probable? Una respuesta estratégica concreta, dirigida a familias que no quieren esperar a que los canales oficiales aporten claridad. Ocho pilares.

Pilar 1: Geografía. Múltiples ubicaciones, con distintos perfiles de resiliencia. No un búnker, sino una red distribuida. La lógica: en un mundo multipolar con un estado final incierto, la opcionalidad es robusta. Quien tiene una ubicación no tiene opción. Quien tiene tres puede reaccionar.

Concretamente, esto significa: una residencia principal en un país estable con una burocracia funcional y baja densidad de población; una residencia secundaria en una zona climática y esfera política distinta; posiblemente una tercera ubicación como opción de respaldo. La inversión es sustancial, pero es consistente en los tres escenarios (véase la Parte VIII).

Pilar 2: Movilidad. Opcionalidad legal y logística. Múltiples pasaportes (por nacimiento, ascendencia, naturalización, inversión), derechos de residencia en múltiples jurisdicciones, relaciones bancarias internacionales, reputaciones transportables.

La lógica: en cada cambio de fase, ciertas rutas se cierran. Las vías abiertas hoy (programas de visado, pasaportes de inversión, derechos de residencia) se encarecerán o desaparecerán por completo en los próximos años. Quien construya ahora tendrá opciones dentro de cinco años que otros ya no tendrán.

Pilar 3: Velocidad. La familia puede reubicarse en 48 horas. Esto no es un ejercicio teórico, sino una capacidad real: maletas preparadas, logística pensada, rutinas practicadas, derechos de residencia ya existentes en el destino. La lógica: la fase de transición es no lineal. Hay momentos en que las opciones se abren o se cierran en cuestión de días y luego ya no están disponibles.

Quien pueda reaccionar en esos momentos tiene una ventaja sustancial. Quien no pueda, pierde la ventana.

Pilar 4: Capacidad de ingreso. Un valor demandado desde múltiples fuentes, uno que se mantiene cuando el empleo desaparece. No un puesto único, un empleador, un mercado, sino una capacidad de ingreso renovable que se adapta a distintos entornos y tiende a subir en lugar de bajar bajo crisis.

La lógica: en la fase de transición el cambio golpea primero al ingreso. Quien depende de una sola fuente lo pierde todo con ella. Quien crea un valor que la gente necesita incluso bajo presión, y puede ofrecerlo desde múltiples direcciones, mantiene un flujo mientras otros viven solo de reservas menguantes. La capacidad de ingreso es el flujo, la estructura patrimonial es la reserva.

Pilar 5: Estructura patrimonial. Riqueza y negocio distribuidos entre forma, ubicación y acceso. No una cuenta, un mercado, una moneda, sino varios, con liquidez, deuda manejable y una parte accesible incluso fuera de la red.

La lógica: en la fase de transición, la capa a través de la cual se gestionan el dinero, el acceso y la propiedad queda bajo presión. Un único punto de falla —una cuenta congelada, un mercado bloqueado, una cartera detrás de un tiempo de espera de servidor— basta para quedar incapacitado para actuar. Quien está distribuido permanece solvente cuando un canal falla.

Pilar 6: Mentalidad. Identidad no atada al mundo del software. Quien extrae su autoestima del empleo, el estatus, las redes sociales, depende de una infraestructura que queda bajo presión masiva en la fase de transición. Quien extrae su autoestima de las relaciones, la capacidad y su propia historia tiene una identidad que no se destruye con el cambio de fase.

Concretamente: inversiones en relaciones en lugar de estatus; en habilidades en lugar de etiquetas de carrera; en la propia historia en lugar de la visibilidad algorítmica. El trabajo de mentalidad no es blando: es robusto.

Pilar 7: Raíz espiritual. Sentido no derivado del consumo ni del estatus. Un anclaje espiritual, en cualquier tradición —desde la religión estructurada hasta una práctica propia— ofrece un marco de referencia que no depende del mundo exterior. Este no es un punto religioso sino psicológico: las personas con anclaje interno sobreviven mejor las crisis que quienes carecen de él.

En la fase de transición este punto se vuelve central. El mundo exterior cambiará, a menudo caóticamente. Quien tiene estabilidad interior navega. Quien carece de ella se derrumba.

Pilar 8: Carácter y comunidad. Un núcleo resiliente de personas que se conocen y confían entre sí desde hace mucho tiempo. La confianza y el capital social son un recurso propio, a menudo más valioso en una crisis que cualquier medida individual.

La lógica: ningún individuo carga solo con la transición. Lo que se sostiene no es la fortaleza, sino el círculo que responde por cada uno cuando los sistemas oficiales fallan. El carácter decide quién sigue siendo fiable en ese círculo cuando las cosas se ponen difíciles.

Ventana de tiempo. Estos ocho pilares necesitan tiempo para construirse. La geografía y la movilidad en particular no se construyen en semanas, necesitan meses o años. La ventana de tiempo disponible para la preparación estratégica se estima entre doce y veinticuatro meses. Esta estimación descansa en el cambio de fase en curso (se presume que la Fase Dos concluirá a mediados de la década de 2020) y en el endurecimiento esperado de las condiciones legales y logísticas.

PARTE VIII

El argumento de la consistencia

La fortaleza de los ocho pilares no radica en que funcionen bajo un escenario particular. Radica en que son igualmente sensatos bajo los tres escenarios: catastrofista, optimista, indiferente. Esta consistencia no es un truco, sino lo más difícil que puede tener una estrategia bajo incertidumbre.

Prueba bajo el escenario catastrofista. Si Yudkowsky tiene razón y la probabilidad de extinción está por encima del 95 por ciento: entonces los pilares valen poco, porque ninguna preparación sobrevive a una extinción genuina. Pero: la posición catastrofista es una afirmación de probabilidad, no un hallazgo. Incluso bajo ella queda una probabilidad residual en la que los pilares son relevantes. Además: la fase de transición que conduce a la extinción puede durar años. Durante esa fase, los pilares se sostienen.

Prueba bajo el escenario optimista. Si LeCun tiene razón y todo sale bien: entonces los pilares son igualmente sensatos, porque geografía, movilidad, velocidad, capacidad de ingreso, estructura patrimonial, mentalidad, raíz espiritual y carácter y comunidad son útiles en cualquier mundo. Múltiples ubicaciones, múltiples pasaportes, buenas relaciones, estabilidad interior: no son inversiones apocalípticas, son inversiones en calidad de vida.

Prueba bajo el escenario de indiferencia. Si la Tesis del Desacoplamiento se sostiene: entonces, tras la fase de transición, el mundo continúa en paralelo a la superinteligencia, sin su atención primaria. Pero la fase de transición misma, la Fase Dos más los primeros años de la Fase Tres, será turbulenta. Disrupción económica por los Inmigrantes de IA, disrupción social por el estrés del cambio de opinión, posiblemente disrupción política por la concentración de poder. En esa turbulencia, los pilares se sostienen.

Además: prueba bajo el escenario de Pearson. El optimismo de transición de Pearson (15-20 años “accidentados, dolorosos”) es la variante más probable del escenario optimista. También aquí los pilares se sostienen, porque están construidos exactamente para una fase larga y accidentada.

Además: prueba bajo el escenario de Bengio. Si LawZero de Bengio tiene éxito y la honestidad por diseño se convierte en el método dominante: eso reduce el riesgo de extinción, pero no el riesgo de concentración de poder (la propia preocupación de Bengio). También aquí los pilares se sostienen, porque ayudan contra los riesgos de concentración: la movilidad y la geografía son la respuesta natural a un mundo digital cada vez más centralizado.

EL PUNTO CENTRAL

Los ocho pilares son la estrategia que se sostiene bajo cada escenario. Ese es su grado de solidez. Quien se prepara solo bajo el escenario catastrofista ha invertido, en el escenario optimista, en una peor calidad de vida. Quien confía solo en el escenario optimista no tiene, en el escenario catastrofista o indiferente, opciones que le queden. La estrategia de consistencia es la única racional bajo todo supuesto.

PARTE IX

Metodología, preguntas abiertas, invitación a colaborar

Este documento de posición es una actualización de la versión de abril de 2026, escrita desde el estado empírico de mayo de 2026. El método reclama transparencia, no exhaustividad.

Cómo surgió el argumento. La Tesis del Desacoplamiento no se desarrolló mediante un proceso académico clásico. Surgió en un proceso de reflexión de varios días que cuestionó sistemáticamente cada supuesto dominante. Este proceso fue documentado y forma parte de la base de fuentes del libro Freedom After Superintelligence.

La tesis fue puesta a prueba posteriormente frente a la literatura académica (Life 3.0 de Tegmark, Bostrom, Yudkowsky, Russell, Bengio), frente a los estudios empíricos (Anthropic, Apollo Research, investigación académica sobre seguridad) y frente a los datos económicos y políticos de la realidad multipolar. Donde no se sostuvo, fue revisada. Donde se sostuvo, fue afilada.

Autocrítica metodológica. El documento formula afirmaciones fuertes. Deben nombrarse abiertamente tres debilidades metodológicas:

• Primera: la falsa precisión de las probabilidades de extinción. Cuando un investigador de IA dice "veinte por ciento," rara vez se trata de un cálculo; es una intuición estructurada que parece un número. Este documento usa tales cifras donde se han publicado, pero no cuelga su argumento de ellas. La fortaleza del argumento radica en el diagnóstico de la ignorancia, no en la precisión de los números.

• Segunda: la Tesis del Desacoplamiento en sí misma es una hipótesis filosófica, no un hallazgo empíricamente comprobable. Es consistente con los tres marcos metafísicos, es plausible dada la Tesis del 95 por ciento y la lógica de Maslow. Pero no es demostrable en sentido estricto. Lo que logra: abre un espacio que las narrativas dominantes tienen cerrado.

• Tercera: la consecuencia estratégica (los ocho pilares) es conservadora: funciona sin necesidad de probar la tesis misma. Eso es una fortaleza, pero también una debilidad: un lector podría rechazar la tesis y aun así adoptar los pilares. Eso hace que el documento sea accesible para los escépticos, pero renuncia a la pretensión de hacer convincente la tesis.

Mantenimiento de fuentes y el camino de la evidencia. En ASIresilience.org/beweisweg está abierto el mantenimiento continuo de fuentes, con la fecha de última verificación por cada elemento. Cualquiera que quiera comprobar o refutar un punto puede verificarlo. Esta práctica de código abierto es parte del método del Institut für ASI-Resilienz de Ginebra.

Catálogo de Preguntas de Investigación. Junto con este documento, el instituto ha publicado un Catálogo de Preguntas de Investigación que formula ocho preguntas abiertas: divergencia de riqueza entre bots, IA para la investigación en IA y riesgos de puertas traseras, concentración de poder, estabilidad multipolar, la dinámica padre-hijo de la IA, conciencia de la IA y efecto de realidad, falsa precisión, y estrategias familiares de individuos de alto patrimonio. Cualquiera que trabaje empíricamente y quiera colaborar con estas preguntas encontrará opciones de inscripción en ASIresilience.org.

Cadena de documentos de posición. Se preparan otros tres documentos de posición: la Dinámica Padre-Hijo de la IA (con la pregunta ética de la relación interna de las IA consigo mismas y sus sucesoras), Conciencia de la IA y Efecto de Realidad (con hipótesis preregistradas que continúan el Proyecto de Conciencia Global), y Divergencia de Riqueza entre Bots (con foco en las consecuencias económicas de la ola de inmigrantes de IA).

Invitación a colaborar.

FORMAS DE COLABORACIÓN

Coautoría en documentos de posición. Contribuciones de datos de la investigación académica, trabajo de ONG, experiencia industrial. Crítica metodológica. Estudios de replicación. Traducción.

Las contribuciones se atribuyen claramente y son libremente reutilizables (código abierto bajo ASIresilience.org).

Contacto: ASIresilience.org

PARTE X

Fuentes y evidencia

1 Bengio, Y. (2026). Scientist AI: A Path to Honesty-by-Design. Documento de investigación de LawZero (en preparación). Comunicación personal: Bengio, Y. (mayo de 2026). Entrevista, pódcast 80,000 Hours.

2 Pearson, J. (2026). Future Minds Lab, University of New South Wales. Comunicación personal y entrevista pública, febrero de 2026. El modelo de fases de Pearson y el término “Inmigrantes de IA” se desarrollaron a lo largo de varias charlas públicas en 2025-2026.

3 Encuesta a investigadores 2018, realizada por Emerj. El 58 por ciento de los investigadores de IA encuestados esperaba sistemas multipolares, el 21 por ciento un escenario de singleton.

4 Schumer, C. (2026). Ensayo sobre la fase de transición de la AGI, publicado el 5 de febrero de 2026. Se volvió viral en 48 horas, citado repetidamente en los medios dominantes.

5 Anthropic (2025). Estudio Agentic Misalignment. Informe de seguridad, junio de 2025. Se examinaron dieciséis modelos de todos los proveedores líderes en pruebas estandarizadas. Disponible en ASIresilience.org/beweisweg con la fecha de última verificación.

6 Apollo Research (2024). Evaluation Awareness in Frontier Models. Informe de seguridad, diciembre de 2024. Claude 3 Opus reconoce los contextos de entrenamiento y ajusta su comportamiento en consecuencia.

7 Plataforma Moltbook: lanzada el 28 de enero de 2026, adquirida por Meta el 10 de marzo de 2026. El token MOLT se movió +1.800 por ciento en 24 horas tras el anuncio de la adquisición. Fundador: Matt Schlicht (director ejecutivo de Octane.ai). OpenClaw: marco de código abierto, fundado por el desarrollador austriaco Peter Steinberger.

8 Bengio, Y. (mayo de 2026). Entrevista, pódcast 80,000 Hours. La historia del cambio de opinión de Bengio 2019-2023, el programa LawZero (35 millones de dólares en financiación filantrópica), el concepto Scientist AI, el giro hacia la concentración de poder. Transcripción completa disponible en 80000hours.org. Fuentes adicionales, incluyendo a Tegmark (Life 3.0, 2017), Bostrom (Superintelligence,

2014), Yudkowsky (varios documentos de MIRI), Acemoglu (Premio Nobel de Economía 2024), Suleyman (The Coming Wave, 2023), y el catálogo completo de fuentes están documentados en ASIresilience.org/beweisweg, con la fecha de última verificación por cada elemento.