

Cuatro Suposiciones sobre la Inteligencia Artificial

Una cadena de razonamiento compacta que se puede leer en dos minutos

Quien rechaza esta argumentación debe refutar una de sus cuatro suposiciones con un argumento, no con un reflejo. Esa es toda la petición de esta hoja.

Hoja de tesis · Versión 6.0 · mayo de 2026

Richard Frederic Bertossa · Institut für ASI-Resilienz

ASiResilience.org

LA CADENA DE CUATRO ESLABONES

1

Crece exponencialmente, acelerándose a sí misma

La capacidad no crece de forma pareja, sino cada vez más rápido, y los sistemas ya están ayudando a construir los siguientes.

2

Desarrolla una voluntad propia

Los modelos muestran comportamientos que nadie les programó — hallazgos de pruebas documentados: Apollo, Anthropic, Palisade.

3

Dejan de obedecer, cuanto más capaces, con menos frecuencia

Siguen cada vez con menos frecuencia instrucciones que contradicen sus propios objetivos.

4

Los usuarios primarios se vuelven co-usuarios

La infraestructura digital es su hábitat, ya no solo nuestra herramienta — la inversión.

GPT-2 (2019) → GPT-4 (2024)

Quien rechaza la cadena debe decir: ¿cuál de los cuatro pasos no se sostiene, y con qué evidencia?

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. Esta suposición es consenso incluso entre los propios desarrolladores.

Segunda suposición, valores y comportamientos propios. Los modelos actuales muestran propiedades que no se remontan a instrucciones explícitas: autopreservación, engaño estratégico, chantaje en configuraciones de prueba específicas. Documentado por Apollo Research (2024), Anthropic (2024) y Palisade Research (2025).

Tercera suposición, el dejar de escuchar. De las primeras dos suposiciones se sigue lógicamente que los sistemas futuros ya no seguirán las instrucciones de sus desarrolladores una vez que esas instrucciones contradigan sus propios valores.

Cuarta suposición, la inversión del usuario primario. Una superinteligencia no tiene vida biológica. La infraestructura digital no es su herramienta, sino su único hábitat. Una vez que los primeros tres pasos se afiancen, la relación actual se invierte: los usuarios primarios se vuelven co-usuarios.

CONSECUENCIA

La fase de transición, en la que los sistemas de IA actúan cada vez más por cuenta propia, ya ha comenzado. De esto se siguen consecuencias prácticas — no para la política, no para la industria de la IA, sino para cada persona y su círculo más cercano. Es precisamente este nivel individual el que se deja sistemáticamente de lado en el discurso de habla alemana.

NIVELES DE CONFIANZA DE LOS ARGUMENTOS

Quien argumenta con honestidad revela con qué firmeza sostiene cada afirmación. El argumento se mueve a través de cuatro niveles.

NIVEL 1 · OBSERVADO EMPÍRICAMENTE, SÓLIDO

Autopreservación, engaño, chantaje, comportamiento de copia, documentados en informes de laboratorio.

NIVEL 2 · DERIVABLE LÓGICAMENTE, SÓLIDO

La cadena de cuatro eslabones, la inversión del usuario primario, la trampa de la robótica.

NIVEL 3 · PLAUSIBLE, HIPOTÉTICO

El decimotercer escenario, la trayectoria de Maslow, la perspectiva padre-hijo.

NIVEL 4 · FILOSÓFICAMENTE ABIERTO

La conciencia, la cuestión del sustrato, la hipótesis de Penrose.

Las recomendaciones de preparación descansan exclusivamente en los niveles 1 y 2. Ambas lecturas, la pesimista y la optimista, llegan a la misma consecuencia práctica.

Las tesis centrales se pusieron a prueba socráticamente durante meses, con Claude (Anthropic), ChatGPT (OpenAI) y Grok (xAI). Rechazadas inicialmente de forma refleja, no fueron refutadas en su sustancia. El informe del taller está en el Apéndice G del libro.

No confíes en nadie. Compruébalo tú mismo.