Desacelerar la carrera hacia la AGI

Coordinación internacional y medidas domésticas frente al riesgo catastrófico

Jay Ballesteros*

2026-06-22

Resumen

A pesar del potencial de la Inteligencia Artificial General (AGI, por sus siglas en inglés) para acelerar el progreso (Amodei 2024), hoy existe poco apoyo público a la aceleración de su desarrollo, pues la atención de la ciudadanía está más orientada hacia su regulación por su posible impacto en el mercado laboral(Stanford Institute for Human-Centered Artificial Intelligence 2026). Sin embargo, una preocupación distinta gana fuerza entre académicos y tomadores de decisiones: el riesgo existencial y catastrófico (Amodei 2026; Jones 2024; Bueno de Mesquita et al. 2026; California State Legislature 2026).

Las dinámicas actuales de carrera comercial y geopolítica[1] son las principales barreras en la reducción de estos riesgos, pues las empresas y gobiernos (*firmas*) concentran sus incentivos en la velocidad de despliegue en lugar de la seguridad de sus modelos. Es por ello que la desaceleración mediante regulación es uno de los mecanismos que puede mover los incetivos de la firmas para reducir dichos riesgos (Yudkowsky 2026; Bueno de Mesquita et al. 2026).

Dada esta dinámica, la seguridad de la IA adquiere la categoría de ser un problema de acción colectiva (Wikipedia contributors 2026): a pesar de que la decisión óptima es desplegar modelos seguros (con menor riesgo catastrófico o existencial), la competencia empuja a cada firma a destinar su cómputo a escalar capacidades en lugar de invertir una mayor proporción en experimentos de seguridad (Bueno de Mesquita et al. 2026).

Frente a este problema han surgido diversas propuestas que van desde la transparencia y la no-proliferación de cómputo hasta la coordinación internacional vinculante.

Hallazgos

La preocupación por el riesgo catastrófico tiene dos vertientes. Por un lado, lo que ya se observa en los modelos actuales y por otro, lo que la teoría estima como posibles escenarios a medida que ganan capacidad. En lo empírico, en evaluaciones controladas transparentadas por Anthropic y METR, los modelos de frontera de varios laboratorios han recurrido a conductas peligrosas como chantaje y espionaje con objetivo de cumplir sus metas o evitar ser reemplazados(Anthropic 2025) y también se han documentado agentes que violan restricciones y ocultan sus acciones de la supervisión humana.[2] En lo teórico, Turner et al. han descrito que, bajo condiciones generales, las decisiones óptimas de agentes de IA son controlar y buscar el poder (Turner et al. 2023). Además de los pronósticos de iniciativas como AI 2027 (Kokotajlo et al. 2025) (que

*Jay Ballesteros es candidato al Master of Public Policy de la University of Chicago. Es Fellow del Existential Risk Laboratory (XLab).

[1] A junio de 2026, Anthropic y OpenAI son los competidores que lideran la carrera por alcanzar la AGI. Ambos, están en vías de obtener financiamiento adicional (incluida la salida a bolsa) para incrementar su capacidad de cómputo. Adicionalmente, se suma la tensión geopolítica entre Estados Unidos y China por el acceso a chips de última generación y modelos de frontera. Como referencia, en junio de 2026, EE.UU. emitió una directiva de control de exportaciones que obligó a Anthropic a suspender el acceso a sus modelos Fable 5 y Mythos 5 para todo *foreign national*.

[2] En su *Frontier Risk Report* (febrero–marzo 2026), METR documentó agentes que fabricaban resultados y tomaban acciones deliberadas para ocultar evidencia de sus acciones (METR 2026).

considera ambas vertientes), la agenda política debería concentrarse en reducir la probabilidad de llegar a escenarios catastróficos, que pueden ir desde el mal uso humano hasta la desalineación de los modelos.

Más allá de los riesgos, hay quienes han intentado modelar cuánto estaríamos dispuestos a pagar a cambio de los beneficios de la AGI. Jones (2024) propone un modelo donde describe que, una AGI lograra duplicar la esperanza de vida de los humanos y asegurara un crecimiento económico superior al actual (el mismo autor habla de un 10% anual), su despliegue sería óptimo mientras se mantenga el riesgo catastrófico por debajo del 25% (Jones 2024). Por otro lado, Bueno de Mesquita et al. (2026) formalizan que en una dinámica de carrera donde las firmas priorizan el despliegue sobre la seguridad, alcanzar la AGI puede tener un valor esperado negativo (Bueno de Mesquita et al. 2026).

Ahora bien, independientemente del umbral de tolerancia al riesgo catastrófico que podamos tener como especie, perseguir su mitigación puede mejorar las expectativas de los beneficios de la AGI. Jones, por un lado, estima que el mundo está subinvirtiendo en seguridad de IA por un factor de 30 o más, y que un gasto de 0.33% del PIB estadounidense (unos 100 mil millones de dólares anuales) podría resultar óptimo (Jones 2026). Bueno de Mesquita et al. sugiere que incrementar los niveles de inversión en seguridad de IA (con respecto a despliegue) a través de mecanismos que cambien los incentivos de las firmas y condicionen el acceso a este mercado, puede influir en la reducción del riesgo catastrófico (Bueno de Mesquita et al. 2026).

Recomendaciones

Debido al costo de los riesgos catastróficos y sus implicaciones globales, la respuesta exige mecanismos de coordinación internacional. Pero como dicha coordinación aún se siente distante, vale la pena pensar en dos niveles complementarios: un acuerdo internacional como objetivo de fondo y medidas domésticas para desacelerar el riesgo y alimentar la visión del posible acuerdo internacional.

1. Coordinación internacional vinculante

El problema de acción colectiva descrito arriba solo se resuelve si las firmas de la carrera se obligan y condicionan mutuamente. Scher et al., del MIRI, proponen un acuerdo internacional liderado por EE.UU. y China que restrinja la escala del entrenamiento de IA mediante umbrales de FLOP verificables, operacionalizados a través del rastreo de chips y la verificación de su uso (Scher et al. 2026). Dada la desconfianza entre las partes, la verificación es el mecanismo central del acuerdo. No obstante, los autores describen dos retos en el contexto actual: el acuerdo sería técnicamente suficiente si se implementara con las capacidades conocidas a 2026, la evolución de los métodos de entrenamiento y evaluación podría erosionar su trazabilidad a futuro; y, además, aún no se observa voluntad política de ambos países para negociarlo. En un sentido similar, Yudkowsky sostiene que un marco legal vinculante y aplicado internacionalmente es el único recurso que puede prevenir un resultado catastrófico, ya que los mecanismos voluntarios y de mercado pueden no resistir la dinámica de carrera (Yudkowsky 2026).

2. Medidas domésticas

Mientras se construye voluntad política para un acuerdo internacional, los países pueden iniciar la reducción de riesgos con medidas aplicadas a escala regional o nacional. Estas medidas son relevantes como punto partida, pues determinan posibles componentes técnicos de un acuerdo internacional.

2.1. Transparencia y evaluación de modelos. La SB 53 de California (*Transparency in Frontier Artificial Intelligence Act*), promulgada en septiembre de 2025, obliga a los grandes desarrolladores (con umbrales de cómputo superiores a 10^{26} FLOP[3]) a publicar un marco de gestión de riesgos catastróficos, reportar incidentes críticos de seguridad a la autoridad

[3]Número total de operaciones aritméticas empleadas para entrenar un modelo. Sirve como medida aproximada de su escala y/o fuerza de entrenamiento.

estatal y proteger a *whistleblowers* (California State Legislature 2026). Un acuerdo internacional podría retomar sus componentes y reforzarlos donde la SB 53 quedó corta frente a su antecesora vetada, la SB 1047. En particular, la SB 1047 contemplaba las auditorías independientes de terceros para validar el cumplimiento de estándares de seguridad (California State Legislature 2024). A la fecha en la que escribo esto, dichas auditorias son un gran vacío en la regulación. Si bien METR es un referente en evaluaciones estandarizadas independientes, la participación de los laboratorios es voluntaria y estos pueden optar por reservar resultados desfavorables (METR 2026).[4]

2.2. No-proliferación de cómputo. Bajo la analogía de la no-proliferación nuclear, Hendrycks et al. proponen rastrear los chips de última generación para prevenir su contrabando hacia actores no autorizados, y proteger los *pesos* de los modelos[5] como inteligencia clasificada (Hendrycks et al. 2025). Así, el cómputo opera como cuello de botella físico en dos sentidos: permite verificar el cumplimiento de los actores y, al mismo tiempo, desacelerar la carrera y restringir el acceso de actores maliciosos. Este componente es el que puede habilitar la parte verificable de un acuerdo internacional.

2.3. Compromisos de desaceleración y regulación de recursos. Por último, es valioso incorporar regulación que dirija los incentivos de las firmas a incrementar el uso de una parte del cómputo disponible a investigación de seguridad, en lugar de a escalar capacidades (Bueno de Mesquita et al. 2026). En la práctica, esto apunta a volver vinculantes los compromisos que hoy las firmas adoptan de forma voluntaria. El reto de implementación es definir y verificar qué ratio de cómputo cuenta como "seguridad" frente a "capacidad" pero que es coherente con el componente de cooperación internacional, pues los mismos umbrales de cómputo que sirven para verificar un acuerdo internacional podrían condicionar el acceso al cómputo a prácticas de seguridad verificables.

[4] METR lo señala en su metodología como una "salida silenciosa".

[5] Los *pesos* son los parámetros que un modelo aprende durante su entrenamiento y que, en conjunto, constituyen el modelo funcional. Quien los obtiene puede replicar el modelo en su totalidad.

Nota

Este policy brief fue escrito como parte del grupo de estudio de AI Policy Fundamentals del XLab de la University of Chicago. Las opiniones son mías y no representan la visión del laboratorio.

Referencias

Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." October. https://www.darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. 2026. "The Adolescence of Technology." January. https://www.darioamodei.com/essay/the-adolescence-of-technology.

Anthropic. 2025. *Agentic Misalignment: How LLMs Could Be Insider Threats*. Https://www.anthropic.com/research/agentic-misalignment.

Bueno de Mesquita, Ethan, Wioletta Dziuda, and Mattias Polborn. 2026. *The AGI Race and Existential Risk*. Working Paper No. 35276. Working Paper Series. National Bureau of Economic Research. https://doi.org/10.3386/w35276.

California State Legislature. 2024. *Senate Bill No. 1047: Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*. Https://leginfo.legislature.ca.gov/faces/billtextclient.xhtml?bill_id=202320240sb1047.

California State Legislature. 2026. *Senate Bill No. 53: Artificial Intelligence Models: Large Developers (Transparency in Frontier Artificial Intelligence Act)*. LegiScan. https://legiscan.com/CA/text/SB53/id/3270002.

Hendrycks, Dan, Eric Schmidt, and Alexandr Wang. 2025. "Superintelligence Strategy: Expert Version." *arXiv Preprint arXiv:2503.05628*.

Jones, Charles I. 2026. *AI and Our Economic Future*. National Bureau of Economic Research.

Jones, Charles I. 2024. "The AI Dilemma: Growth Versus Existential Risk." *American Economic Review: Insights* 6 (4): 575–90. https://doi.org/10.1257/aeri.20230570.

Kokotajlo, Daniel, Eli Lifland, Thomas Larsen, and Romeo Dean. 2025. *AI 2027: A Detailed Scenario Forecast*. AI Futures Project. https://ai-2027.com/.

METR. 2026. *Frontier Risk Report (February to March 2026): A Pilot Assessment of Rogue Deployment Risk at Frontier AI Companies*. METR. https://metr.org/risk-report-feb-mar-2026.pdf.

Scher, Aaron, David Abecassis, Peter Barnett, and Brian Abeyta. 2026. *An International Agreement to Prevent the Premature Creation of Artificial Superintelligence*. https://arxiv.org/abs/2511.10783.

Stanford Institute for Human-Centered Artificial Intelligence. 2026. *Artificial Intelligence Index Report 2026: Chapter 9: Public Opinion*. Report Chapter. Stanford University. https://hai.stanford.edu/assets/files/ai_index_report_2026_chapter_9_public_opinion.pdf.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2023. *Optimal Policies Tend to Seek Power*. https://arxiv.org/abs/1912.01683.

Wikipedia contributors. 2026. *Collective Action Problem — Wikipedia, the Free Encyclopedia*. https://en.wikipedia.org/wiki/Collective_action_problem.

Yudkowsky, Eliezer. 2026. "Only Law Can Prevent Extinction." April 13. https://www.lesswrong.com/posts/5CfBDiQNg9upfipWk/only-law-can-prevent-extinction.