AI ACTION SUMMIT

Informe internacional sobre la seguridad de la IA

Informe científico internacional
sobre la seguridad de la IA avanzada

Enero de 2025

Colaboradores

PRESIDENTE

Prof. Yoshua Bengio, Universidad de Montreal/Mila – Instituto de IA de Quebec

PANEL ASESOR DE EXPERTOS

Este panel internacional fue nominado por los gobiernos de los 30 países enumerados a continuación, la ONU, la UE y la OCDE.

Alemania: Antonio Krüger, Centro Alemán de Investigación en Inteligencia Artificial

Australia: Bronwyn Fox, Universidad de Nueva Gales del Sur

Brasil: André Carlos Ponce de Leon Ferreira de Carvalho, Instituto de Matemáticas e Informática, Universidad de São Paulo

Canadá: Mona Nemer, Asesora científica principal de Canadá

Chile: Raquel Pezoa Rivera, Universidad Técnica Federico Santa María

China: Yi Zeng, Academia de Ciencias de China

Emiratos Árabes Unidos: Marwan Alserkal, Ministerio de Asuntos del Gabinete, Oficina del Primer Ministro

España: Nuria Oliver, ELLIS Alicante

Estados Unidos: El Saif M. Khan, Ministerio de Comercio de EE. UU.

Filipinas: Dominic Vincent Ligot, CirroLytix

Francia: Guillaume Avrin, Coordinador Nacional de Inteligencia Artificial

India: Balaraman Ravindran, Escuela Wadhwani de Ciencias de Datos e Inteligencia Artificial, Instituto Indio de Tecnología de Madrás

Indonesia: Hammam Riza, Investigación Colaborativa e Innovación Industrial en Inteligencia Artificial (KORIKA)

Irlanda: Ciarán Seoighe, Research Ireland

Israel: Ziv Katzir, Autoridad de Innovación de Israel

Italia: Andrea Monti, Experto jurídico de la Subsecretaría de Estado para la Transformación Digital, Presidencia del Consejo de Ministros Italiano

Japón: Hiroaki Kitano, Sony Group Corporation

Kenia: Nusu Mwamanzi, Ministerio de TIC y Economía Digital

México: José Ramón López Portillo, LobsterTel

Naciones Unidas: Amandeep Singh Gill, Secretario General Adjunto de Tecnologías Digitales y Emergentes y Enviado del Secretario General para la Tecnología

Nigeria: Olubunmi Ajala, Ministerio de Comunicaciones, Innovación y Economía Digital

Nueva Zelanda: Gill Jolly, Ministerio de Negocios, Innovación y Empleo

OCDE: Jerry Sheehan, Director de la Dirección de Ciencia, Tecnología e Innovación

Países Bajos: Haroon Sheikh, Consejo Científico de Políticas Gubernamentales de los Países Bajos

República de Corea: Kyoung Mu Lee, Departamento de Ingeniería Eléctrica e Informática, Universidad Nacional de Seúl

Ruanda: Crystal Rugege, Centro para la Cuarta Revolución Industrial

Singapur: Denise Wong, Grupo de Innovación y Protección de Datos, Autoridad de Desarrollo de Medios de Infocomm

Suiza: Christian Busch, Departamento Federal de Asuntos Económicos, Educación e Investigación

Turquía: Ahmet Halit Hatip, Ministerio de Industria y Tecnología de Turquía

Ucrania: Oleksii Molchanovskyi, Comité de Expertos sobre el Desarrollo de la Inteligencia Artificial en Ucrania

Unión Europea: Juha Heikkilä, Oficina Europea de IA

Reino de Arabia Saudita: Fahad Albalawi, Autoridad Saudita de Datos e Inteligencia Artificial

Reino Unido: Chris Johnson, Asesor científico jefe del Departamento de Ciencia, Innovación y Tecnología

DIRECTOR CIENTÍFICO

Sören Mindermann, Mila – Instituto de IA de Quebec

ESCRITOR PRINCIPAL

Daniel Privitera, Centro KIRA

GRUPO DE ESCRITURA

Tamay Besiroglu, Epoch AI

Rishi Bommasani, Universidad de Stanford

Stephen Casper, Instituto Tecnológico de Massachusetts

Yejin Choi, Universidad de Stanford

Philip Fox, Centro KIRA

Ben Garfinkel, Universidad de Oxford

Danielle Goldfarb, Mila – Instituto de Inteligencia Artificial de Quebec

Hoda Heidari, Universidad Carnegie Mellon

Anson Ho, Epoch AI

Sayash Kapoor, Universidad de Princeton

Leila Khalatbari, Universidad de Ciencia y Tecnología de Hong Kong

Shayne Longpre, Instituto Tecnológico de Massachusetts

Sam Manning, Centro para la Gobernanza de la IA

Vasilios Mavroudis, Instituto Alan Turing

Mantas Mazeika, Universidad de Illinois en Urbana-Champaign

Julian Michael, Universidad de Nueva York

Jessica Newman, Universidad de California, Berkeley

Kwan Yee Ng, Concordia AI

Chinasa T. Okolo, Institución Brookings

Deborah Raji, Universidad de California, Berkeley

Girish Sastry, independiente

ASESORES SÉNIOR

Daron Acemoglu, Instituto Tecnológico de Massachusetts

Olubayo Adekanmbi, contribuyó como asesor sénior antes de asumir su cargo en EqualyzAI

David Dalrymple, Agencia de Investigación e Invención Avanzada

Tomás G. Dietterich, Universidad Estatal de Oregón

Eduardo W. Fieltro, Universidad de Princeton

Pascale Fung, contribuyó como asesora sénior antes de asumir su cargo en Meta

Pierre-Olivier Gourinchas, Departamento de Investigación, Fondo Monetario Internacional

Fredrik Heintz, Universidad de Linköping

Geoffrey Hinton, Universidad de Toronto

Nick Jennings, Universidad de Loughborough

Andreas Krause, ETH Zurich

Susan Leavy, University College Dublin

Percy Liang, Universidad de Stanford

Teresa Ludermir, Universidad Federal de Pernambuco

Vidushi Marda, AI Collaborative

Elizabeth Seger (escritora generalista), Demos

Theodora Skeadas, Humane Intelligence

Tobin South, Instituto Tecnológico de Massachusetts

Emma Strubell, Universidad Carnegie Mellon

Florian Tramèr, ETH Zurich

Lucía Velasco, Universidad de Maastricht

Nicole Wheeler, Universidad de Birmingham

Helen Margetts, Universidad de Oxford

John McDermid, Universidad de York

Jane Munga, Fundación Carnegie para la Paz Internacional

Arvind Narayanan, Universidad de Princeton

Alondra Nelson, Instituto de Estudios Avanzados

Clara Neppel, IEEE

Alice Oh, Escuela de Informática KAIST

Gopal Ramchurn, responsable de IA en el Reino Unido

Stuart Russell, Universidad de California, Berkeley

Marietje Schaake, Universidad de Stanford

Bernhard Schölkopf, ELLIS Institute Tübingen

Dawn Song, Universidad de California, Berkeley

Alvaro Soto, Pontificia Universidad Católica de Chile

Lee Tiedrich, Universidad de Duke

Gaël Varoquaux, Inria

Andrew Yao, Instituto de Ciencias de la Información Interdisciplinaria, Universidad de Tsinghua

Ya-Qin Zhang, Universidad de Tsing

SECRETARÍA

Instituto de Seguridad de IA

Baran Acar

Ben Clifford

Rian Overy

Ben Snodin

Mila — Instituto de Inteligencia Artificial de Quebec

Lambrini Das

Claire Dennis

Freya Hempleman

Hannah Merchant

Jonathan Barry

Benjamin Prud'homme

AGRADECIMIENTOS

Revisores de la sociedad civil y la industria

Sociedad civil: Instituto Ada Lovelace, AI Forum New Zealand/Te Kāhui Atamai Iahiko o Aotearoa, Grupo Temporal de Expertos en IA de Australia, Fundación Carnegie para la Paz Internacional, Centro de Derecho e Innovación/Fundación Certa, Centro para la Gobernanza de la IA, Centro de Derecho e Innovación Digital del Presidente de la Corte Suprema Meir Shamgar, Eon Institute, Gradient Institute, Israel Democracy Institute, Fundación Mozilla, Old Ways New, RAND, SaferAI, El Centro para la Resiliencia a Largo Plazo, The Future Society, The Alan Turing Institute, The Royal Society, Asociación de Políticas de Inteligencia Artificial de Turquía.

Industria: Advai, Anthropic, Cohere, Deloitte Consulting USA y Deloitte LLM UK, G42, Google DeepMind, Harmony Intelligence, Hugging Face, IBM, Lelapa AI, Meta, Microsoft, Shutterstock, Zhipu.ai.

Agradecimientos especiales

La Secretaría agradece el apoyo, los comentarios y los aportes de Angie Abdilla, Nitarshan Rajkumar, Geoffrey Irving, Shannon Vallor, Rebecca Finlay y Andrew Strait.

Cualquier consulta sobre esta publicación se debe enviar a: secretariat.AIStateofScience@dsit.gov.uk.

Las consultas sobre el contenido del informe también deben enviarse al responsable científico.

Descargo de responsabilidad

El informe no representa las opiniones del Presidente ni de ninguna persona en particular de los grupos de redacción o asesoramiento, ni de ninguno de los gobiernos que han brindado su apoyo en su elaboración. Este informe es una síntesis de la investigación existente sobre las capacidades y los riesgos de la IA avanzada. El Presidente del informe es el máximo responsable de este y ha supervisado su desarrollo de principio a fin.

Número de serie de investigación: DSIT 2025/001

Acerca de este informe

- **Este es el primer Informe internacional sobre la seguridad de la IA.** Tras una publicación provisional en mayo de 2024, un diverso grupo de 96 expertos en inteligencia artificial (IA) contribuyó a este primer informe completo, incluido un Panel asesor de expertos internacionales nominado por 30 países, la Organización para la Cooperación y el Desarrollo Económicos (OCDE), la Unión Europea (UE) y las Naciones Unidas (ONU). El informe pretende proporcionar información científica que respalde la formulación de políticas informadas. No recomienda políticas específicas.
- **El informe es obra de expertos independientes.** Los expertos independientes que redactaron colectivamente este informe, liderados por el Presidente, tuvieron plena discreción sobre su contenido.
- **Si bien este informe se centra en los riesgos y la seguridad de la IA, la IA también ofrece muchos beneficios potenciales para las personas, las empresas y la sociedad.** Existen muchos tipos de IA, cada uno con diferentes beneficios y riesgos. La mayoría de las veces, y en la mayoría de las aplicaciones, la IA aumenta la eficacia de las personas y las organizaciones. Pero el mundo sólo podrá disfrutar plenamente y de forma segura de los numerosos beneficios potenciales de la IA si sus riesgos se gestionan adecuadamente. Este informe se centra en identificar dichos riesgos y evaluar métodos para mitigarlos. No tiene como fin evaluar exhaustivamente todos los posibles impactos sociales de la IA, ni sus numerosos beneficios potenciales.
- **El informe se centra en la IA de propósito general.** El informe se limita a un tipo de IA que ha avanzado con especial rapidez en los últimos años y cuyos riesgos asociados han sido menos estudiados y comprendidos: la IA de propósito general o la IA que puede realizar una amplia variedad de tareas. El análisis de este informe se centra en los sistemas de IA de propósito general más avanzados al momento de redactar este informe, así como en sistemas futuros que podrían ser aún más capaces.
- **El informe resume la evidencia científica en tres preguntas fundamentales:** ¿Qué puede hacer la IA de propósito general? ¿Cuáles son los riesgos asociados con la IA de propósito general? ¿Y qué técnicas de mitigación existen frente a estos riesgos?
- **Es mucho lo que está en juego.** Nosotros, los expertos que contribuimos a este informe, seguimos en desacuerdo sobre varias cuestiones, menores y mayores, sobre las capacidades de la IA de propósito general, los riesgos y la mitigación de riesgos. Sin embargo, consideramos que este informe es esencial para mejorar nuestra comprensión colectiva de esta tecnología y sus riesgos potenciales. Esperamos que el informe ayude a la comunidad internacional a avanzar hacia un mayor consenso sobre la IA de propósito general y a mitigar sus riesgos de manera más efectiva, para que las personas puedan experimentar sus numerosos beneficios potenciales de manera segura . Es mucho lo que está en juego. Esperamos continuar con esta iniciativa.

Actualización sobre los últimos avances en IA después de la redacción de este informe: Nota del P residente

Desde que terminó la redacción de este informe (5 de diciembre de 2024) hasta su publicación en enero de 2025, ocurrió un acontecimiento importante. La empresa de inteligencia artificial OpenAI compartió los primeros resultados de las pruebas de un nuevo modelo de IA, o3. Estos resultados indican un rendimiento considerablemente superior que cualquier modelo anterior en varias de las pruebas más desafiantes del campo de programación, razonamiento abstracto y razonamiento científico. En algunas de estas pruebas, o3 supera a muchos expertos humanos (pero no a todos). Además, logra un gran avance en una prueba clave de razonamiento abstracto que muchos expertos, incluido yo mismo, creíamos que estaba fuera del alcance del sistema de IA hasta hace poco. Sin embargo, al momento de la redacción de este artículo, no hay información pública sobre sus capacidades en el mundo real, en particular, para resolver tareas más abiertas.

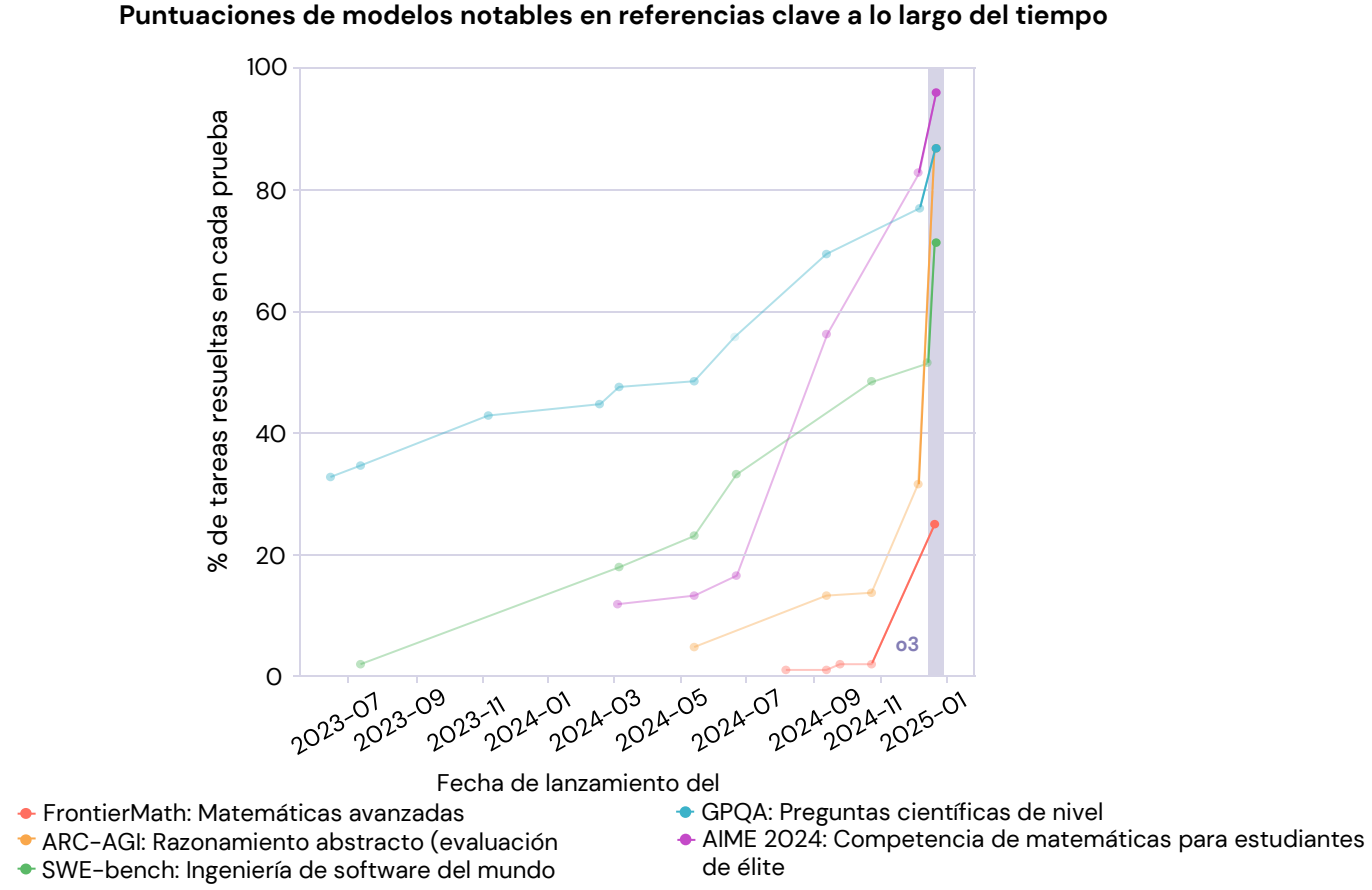

Figura 0.1: *Puntuaciones de modelos de IA de propósito general notables en referencias clave desde junio de 2023 hasta diciembre de 2024. o3 demostró que ha mejorado significativamente su rendimiento en comparación con el anterior (área sombreada). Estas referencias son algunas de las pruebas más desafiantes del campo de la programación, el razonamiento abstracto y el razonamiento científico. Para el modelo o3 que no se lanzó, se muestra la fecha del anuncio; para los otros modelos, se muestra la fecha de lanzamiento. Algunos de los modelos de IA más recientes, incluido o3, se*

beneficiaron de un andamiaje mejorado y más cálculo durante la prueba. Fuentes: Anthropic, 2024; Chollet, 2024; *Chollet et al., 2025*; *Epoch AI, 2024*; *Glazer et al. 2024*; *OpenAI, 2024a*; *OpenAI, 2024b*; *Jimenez et al., 2024*; *Jimenez et al., 2025.*

Los resultados de o3 evidencian que el ritmo de los avances en las capacidades de la IA puede seguir siendo alto o incluso acelerarse. Más específicamente, sugieren que incrementar la capacidad de cálculo de los modelos para resolver un problema determinado ("escalamiento de inferencia") ayuda a superar las limitaciones anteriores. En términos generales, el escalamiento de inferencia aumenta el costo de uso de los modelos. Sin embargo, como lo demuestra otro modelo notable reciente, R1, lanzado por la empresa DeepSeek en enero de 2025, investigadores están logrando reducir estos costos. En general, el escalamiento de inferencia podría permitir a los desarrolladores de IA lograr más avances en el futuro. Los resultados de o3 también destacan la necesidad de comprender mejor cómo el uso creciente de la IA por parte de los desarrolladores de IA puede afectar la velocidad del desarrollo futuro de la IA en sí.

Las tendencias evidenciadas por o3 podrían tener implicaciones importantes en los riesgos de la IA. Los avances previos en la ciencia y las capacidades de programación generaron mayor evidencia sobre los riesgos como ataques cibernéticos y biológicos. Los resultados de o3 también son relevantes para los posibles impactos en el mercado laboral, el riesgo de pérdida de control y el uso de energía, entre otros. Pero las capacidades de o3 también podrían utilizarse para ayudar a brindar protección contra fallos de funcionamiento y usos maliciosos. En general, las evaluaciones de riesgos de este informe se deben analizar con la comprensión de que la IA ha adquirido capacidades desde que se redactó el informe. Sin embargo, hasta el momento, aún no hay evidencia sobre los impactos reales de o3, ni información para confirmar o descartar riesgos nuevos o inmediatos importantes.

La mejora de las capacidades sugerida por los resultados de o3 y nuestra limitada comprensión de las implicaciones para los riesgos de la IA subrayan un desafío clave para los responsables de formular políticas que este informe identifica: tendrán que sopesar frecuentemente los posibles beneficios y riesgos de los avances inminentes en inteligencia artificial sin tener mucha evidencia científica disponible. No obstante, generar evidencia sobre las implicaciones de seguridad y protección de las tendencias implicadas por o3 será una prioridad urgente para la investigación de IA en las próximas semanas y meses.

Conclusiones clave del informe

- **Las capacidades de la IA de propósito general, el tipo de IA en el que se centra este informe, han aumentado rápidamente en los últimos años y han mejorado aún más en los últimos meses.** Hace unos años, los mejores modelos lingüísticos amplios (MLA) rara vez podían producir un párrafo de texto coherente. En la actualidad, la IA de propósito general puede escribir programas informáticos, generar imágenes fotorrealistas personalizadas y participar en conversaciones abiertas prolongadas. Desde la publicación del Informe Interino (mayo de 2024), los nuevos modelos han demostrado un rendimiento notablemente superior en pruebas de razonamiento científico y programación.
- **Actualmente, muchas empresas están invirtiendo en el desarrollo de agentes de IA de propósito general como una posible vía para seguir avanzando.** Los agentes de IA son sistemas de IA de propósito general que pueden actuar, planificar y delegar de forma autónoma para lograr objetivos con poca o ninguna supervisión humana. Los agentes de IA sofisticados podrían, por ejemplo, usar computadoras para completar proyectos más largos que los sistemas actuales, lo que generaría beneficios y riesgos adicionales.
- **Los avances en materia de capacidad que se producirán en los próximos meses y años podrían ser desde lentos hasta extremadamente rápidos.** El progreso dependerá de la capacidad de las empresas para desplegar rápidamente aún más datos y poder computacional para entrenar nuevos modelos, y de la posibilidad de este "escalamiento" de los modelos para superar sus limitaciones actuales. Investigaciones recientes sugieren que los modelos que escalan rápidamente pueden seguir siendo físicamente factibles durante al menos varios años. Pero para lograr avances importantes en materia de capacidad, es posible que también se necesiten otros factores: por ejemplo, nuevos avances en materia de investigación, los cuales son difíciles de predecir, o el éxito de un nuevo enfoque de escalamiento que las empresas hayan adoptado recientemente.
- **Ya están bien establecidos varios de los daños que puede causar la IA de propósito general.** Estos incluyen estafas, imágenes íntimas no consensuadas (IINC) y material de abuso sexual infantil (MASI), resultados sesgados de modelos contra ciertos grupos de personas o ciertas opiniones, problemas de confiabilidad y violaciones de la privacidad. Investigadores han desarrollado técnicas de mitigación para estos problemas, pero hasta ahora ninguna combinación de técnicas puede resolverlos completamente. Desde la publicación del Informe Interino se han revelado formas más sutiles de sesgo a través de nueva evidencia de discriminación relacionada con los sistemas de IA de propósito general.
- **A medida que aumentan las capacidades de la IA de propósito general, surgen gradualmente evidencias de riesgos adicionales.** Estos incluyen riesgos como impactos a gran escala en el mercado laboral, ataques biológicos o hackeos realizados con IA, y la pérdida del control de la sociedad sobre la IA de propósito general. La interpretación de esta evidencia difiere entre los expertos:

Consulte la actualización del presidente sobre los últimos avances en IA después de la redacción de este informe.

mientras algunos piensan que esos riesgos están a décadas de distancia, otros piensan que la IA de propósito general podría producir daños a escala social en los próximos años. Los avances recientes en las capacidades de la IA de propósito general (particularmente en pruebas de razonamiento científico y programación) han generado nueva evidencia de riesgos potenciales como los hackeos y los ataques biológicos realizados con IA, lo que llevó a que una empresa importante de IA aumentara su evaluación del riesgo biológico de su mejor modelo de "bajo" a "medio".

- **Las técnicas de gestión de riesgos siguen en sus primeras etapas, pero el progreso es posible.** Existen varios métodos técnicos que los desarrolladores pueden emplear y los reguladores pueden exigir para evaluar y reducir los riesgos de la IA de propósito general , pero todos tienen limitaciones. Por ejemplo, las técnicas de interpretación actuales para explicar por qué un modelo de IA de propósito general produjo un resultado determinado siguen siendo sumamente limitadas. Sin embargo, los investigadores están logrando algunos avances para abordar estas limitaciones. Por otro lado, los investigadores y las personas responsables de formular políticas intentan estandarizar cada vez más los enfoques de gestión de riesgos y coordinarlos a nivel internacional.
- **El ritmo y la imprevisibilidad de los avances en la IA de propósito general plantean un "dilema de evidencia" para las personas responsables de formular políticas.** Dados los avances, que a veces son rápidos e inesperados, las personas responsables de formular políticas a menudo tendrán que medir los posibles beneficios y riesgos de los avances inminentes en inteligencia artificial sin tener suficiente evidencia científica disponible. Al hacerlo, se enfrentan a un dilema. Por un lado, las medidas preventivas de mitigación de riesgos basadas en evidencia limitada podrían resultar ineficaces o innecesarias. Por otro lado, el esperar evidencia más contundente de un riesgo inminente podría dejar a la sociedad desprevenida, o incluso hacer imposible la mitigación, por ejemplo, si ocurren pasos agigantados en las capacidades de la IA y los riesgos asociados. Las empresas y los gobiernos están desarrollando sistemas de alerta temprana y marcos de gestión de riesgos que podrían reducir este dilema. Algunas de ellas activan medidas de mitigación específicas cuando surge nueva evidencia de riesgos, mientras que otras requieren que los desarrolladores proporcionen evidencia de seguridad antes de lanzar un nuevo modelo.
- **Existe un amplio consenso entre investigadores sobre la utilidad de avanzar en las siguientes cuestiones:** ¿Qué tan rápido avanzarán las capacidades de la IA de propósito general en los próximos años y cómo pueden los investigadores medir este progreso de manera confiable? ¿Cuáles son los límites de riesgo razonables para activar las mitigaciones? ¿Cuál es la mejor manera de que las personas responsables de formular políticas accedan a información sobre la IA de uso general que sea relevante para la seguridad pública? ¿Cómo pueden los investigadores, las empresas tecnológicas y los gobiernos evaluar de manera confiable los riesgos del desarrollo y el despliegue de la IA de propósito general? ¿Cómo funcionan internamente los modelos de IA de propósito general? ¿Cómo se puede diseñar la IA de propósito general para que se comporte de manera segura?

Consulte la actualización del presidente sobre los últimos avances en IA después de la redacción de este informe.

- **La IA no es algo independiente a nosotros: Las decisiones que toman las personas determinan su futuro**. El futuro de la tecnología de IA de propósito general es incierto, con una amplia gama de trayectorias posibles incluso en el futuro cercano, incluidos resultados muy positivos y muy negativos. Esta incertidumbre puede generar fatalismo y hacer que la IA se sienta fuera de nuestro control. Pero serán las decisiones de las sociedades y los gobiernos sobre cómo navegar esta incertidumbre las que determinarán nuestro camino. Este informe pretende facilitar un debate constructivo sobre estas decisiones, basado en evidencia .

Consulte la actualización del presidente sobre los últimos avances en IA después de la redacción de este informe.

Resumen ejecutivo

El propósito de este informe

Este informe sintetiza el estado de la comprensión científica de la IA de propósito general (IA que puede realizar una amplia variedad de tareas), con un enfoque en la comprensión y la gestión de sus riesgos.

Este informe resume la evidencia científica sobre la seguridad de la IA de propósito general. El propósito de este informe es ayudar a crear una comprensión internacional compartida de los riesgos de la IA avanzada y cómo pueden mitigarse. Para lograrlo, se centra en la IA que este tipo de IA ha progresado particularmente rápido en los últimos años y se ha implementado extensamente en empresas de tecnología para una variedad de propósitos comerciales y de negocios. El informe sintetiza el estado de la comprensión científica de la IA de propósito general, con un enfoque en la comprensión y gestión de sus riesgos.

En un contexto de rápidos avances, la investigación sobre la IA de propósito general se encuentra actualmente en una época de descubrimientos científicos y, en muchos casos, aún no es una ciencia consolidada. El informe proporciona un resumen de la comprensión científica actual de la IA de propósito general y sus riesgos. Esto incluye la identificación de áreas de consenso científico y áreas donde hay diferentes puntos de vista o brechas en la comprensión científica actual.

El mundo solo podrá disfrutar plenamente y de forma segura de los beneficios potenciales de la IA de propósito general si sus riesgos se gestionan adecuadamente. Este informe se centra en identificar dichos riesgos y evaluar métodos técnicos para analizarlos y mitigarlos, incluidas las formas en que la propia IA de propósito general puede utilizarse para mitigar los riesgos. No tiene como fin evaluar exhaustivamente todos los posibles impactos sociales de la IA de propósito general. Cabe recalcar que los beneficios actuales y futuros de la IA de propósito general, aunque son muy amplios, quedan fuera del alcance de este informe. La formulación de políticas holísticas requiere considerar tanto los beneficios potenciales de la IA de propósito general como los riesgos abordados en este informe. También es necesario tener en cuenta que otros tipos de IA tienen perfiles de riesgo/beneficio diferentes en comparación con la IA de propósito general actual.

Las tres secciones principales del informe resumen la evidencia científica sobre tres preguntas fundamentales: ¿Qué puede hacer la IA de propósito general? ¿Cuáles son los riesgos asociados con la IA de propósito general? Y ¿qué técnicas de mitigación existen contra estos riesgos?

Sección 1 – Capacidades de la IA de propósito general: ¿Qué puede hacer la IA de propósito general ahora y en el futuro?

Las capacidades de la IA de propósito general han mejorado rápidamente en los últimos años y los avances futuros podrían ser desde lentos hasta extremadamente rápidos.

Los riesgos de la IA de capacidad general dependen directamente de sus capacidades y, según muchas métricas, estas últimas han progresado rápidamente. Hace cinco años, los principales modelos de lenguaje de IA de propósito general rara vez podían crear un párrafo de texto coherente. En la actualidad, algunos modelos de IA de propósito general pueden tener conversaciones sobre una amplia gama de temas, escribir programas informáticos o generar videos cortos realistas a partir de una descripción. Sin embargo, resulta técnicamente difícil estimar y describir de manera confiable las capacidades de la IA de propósito general.

Los desarrolladores han aumentado continuamente los recursos utilizados para entrenar nuevos modelos (esto suele llamarse "escalamiento") y perfeccionado los enfoques existentes para utilizar esos recursos de manera más eficiente. Por ejemplo, según estimaciones recientes, los modelos de IA más avanzados han experimentado incrementos anuales de aproximadamente 4 veces en recursos informáticos utilizados para el entrenamiento y de 2,5 veces en el tamaño del conjunto de datos de entrenamiento.

El ritmo del progreso futuro en las capacidades de la IA de propósito general tiene implicaciones sustanciales para gestionar los riesgos emergentes, pero los expertos no se ponen de acuerdo sobre qué esperar incluso en los próximos meses y años. Según las diversas opiniones de los expertos, las capacidades de la IA de propósito general pueden avanzar de forma lenta, rápida o extremadamente rápida.

Los expertos no se ponen de acuerdo sobre el ritmo del progreso futuro debido a las diferentes opiniones sobre la promesa de un mayor "escalamiento", y las empresas están explorando un tipo de escalamiento adicional y nuevo que podría acelerar aún más las capacidades. Si bien el escalamiento suele superar las limitaciones de los sistemas anteriores, los expertos no están de acuerdo sobre su potencial para resolver las limitaciones restantes de los sistemas actuales, como la falta de confiabilidad para actuar en el mundo físico y para ejecutar tareas extendidas en computadoras. En los últimos meses, un nuevo tipo de escalamiento ha mostrado potencial para mejorar aún más las capacidades: en lugar de un simple escalamiento de los recursos utilizados para entrenar modelos, las empresas de IA también están cada vez más interesadas en el "escalamiento de inferencia". Es decir, permitir que un modelo ya entrenado use más cálculos para resolver un problema determinado, por ejemplo, para mejorar su propia solución o para escribir las llamadas "cadenas de pensamiento" que dividen el problema en pasos más simples.

Consulte la actualización del presidente sobre los últimos avances en IA después de la redacción de este informe.

Varias empresas líderes que desarrollan IA de propósito general están apostando al "escalamiento" para seguir impulsando mejoras de rendimiento. Si las tendencias recientes se mantienen, para fines de 2026 algunos modelos de IA de propósito general se entrenarán utilizando aproximadamente 100 veces más recursos informáticos que los modelos más intensivos de 2023. Esto llevaría al uso de 10 000 veces más recursos informáticos de entrenamiento para 2030, combinado con algoritmos que logran mayores capacidades para una cantidad dada de cálculo disponible. Además de este potencial escalamiento de los recursos de entrenamiento, las tendencias recientes como el escalamiento de inferencia y el uso de modelos para generar datos de entrenamiento podrían significar que se utilizarán aún más cálculos en general. Sin embargo, existen posibles obstáculos que impiden seguir aumentando rápidamente los datos y recursos informáticos, como la disponibilidad de datos, los chips de IA, el capital y la capacidad energética local. Las empresas que desarrollan IA de propósito general están trabajando para superar estos posibles obstáculos.

Desde la publicación del Informe Interino (mayo de 2024), la IA de propósito general ha alcanzado un rendimiento de nivel experto en algunas pruebas y competiciones de razonamiento científico y programación, y las empresas han estado realizando grandes esfuerzos para desarrollar agentes autónomos de IA. Los avances en ciencia y programación fueron impulsados por técnicas de escalamiento de inferencia, como la escritura de largas "cadenas de pensamiento". Nuevos estudios sugieren que ampliar aún más estos enfoques, por ejemplo, al permitir que los modelos analicen problemas mediante la escritura de cadenas de pensamiento aún más largas que los modelos actuales, podría conducir a mayores avances en dominios donde el razonamiento es más importante, como la ciencia, la ingeniería de software y la planificación. Además de esta tendencia, las empresas están realizando grandes esfuerzos para desarrollar agentes de IA de propósito general más avanzados, capaces de planificar y actuar de forma autónoma para trabajar hacia un objetivo determinado. Por último, el precio de mercado del uso de un sistema de IA de propósito general con un nivel de capacidad determinado ha caído drásticamente, lo que hace que esta tecnología sea más accesible y se utilice más ampliamente.

Este informe se centra principalmente en los aspectos técnicos del progreso de la IA, pero la velocidad a la que avanzará la IA de propósito general no es una cuestión puramente técnica. El ritmo de los avances futuros también dependerá de factores no técnicos, incluidos potencialmente los enfoques que adopten los gobiernos para regular la IA. Este informe no analiza cómo los diferentes enfoques de regulación podrían afectar la velocidad de desarrollo y adopción de la IA de propósito general.

Sección 2 – Riesgos: ¿Cuáles son los riesgos asociados con la IA de propósito general?

Ya se han establecido varios daños que puede causar la IA de propósito general. A medida que aumentan las capacidades de la IA de propósito general, surgen gradualmente evidencias de riesgos adicionales.

Este informe clasifica los riesgos de la IA de propósito general en tres categorías: riesgos del uso malicioso, riesgos derivados del mal funcionamiento y riesgos sistémicos. Cada una de estas categorías contiene riesgos que ya se han materializado, así como riesgos que podrían materializarse en los próximos años.

Riesgos del uso malintencionado: **los actores maliciosos pueden usar la IA de propósito general para causar daño a individuos, organizaciones o a la sociedad.** Las formas de uso malicioso incluyen:

- **Daño a individuos a través de contenido falso:** Actualmente, los actores maliciosos pueden usar la IA de propósito general para generar contenido falso que dañe a personas de forma específica. Estos usos maliciosos incluyen pornografía "deepfake" no consentida y MASI generado por IA, fraude financiero mediante suplantación de voz, chantaje para extorsión, sabotaje de reputación personal y profesional, y abuso psicológico. Sin embargo, si bien los informes de incidentes de daños causados por contenido falso generado por IA son comunes, aún faltan estadísticas confiables sobre la frecuencia de estos incidentes.
- **Manipulación de la opinión pública:** La IA de propósito general facilita la generación de contenido persuasivo a gran escala. Esto puede ayudar a los actores que buscan manipular la opinión pública, por ejemplo, para afectar resultados políticos. Sin embargo, la evidencia sobre la frecuencia y eficiencia de estos esfuerzos sigue siendo limitada. Las contramedidas técnicas, como las marcas de agua en el contenido, a pesar de ser útiles, generalmente pueden ser eludidas por actores moderadamente sofisticados.
- **Delitos cibernéticos:** La IA de propósito general puede facilitar o acelerar la ejecución de ciberataques por parte de actores maliciosos con distintos niveles de habilidad. Los sistemas actuales han demostrado capacidades en tareas de ciberseguridad de complejidad baja y media, y los actores patrocinados por un estado están explorando activamente la IA para inspeccionar sistemas objetivo. Una nueva investigación ha confirmado que las capacidades de la IA de propósito general relacionadas con los delitos cibernéticos están avanzando significativamente, pero aún no está claro si esto afectará el equilibrio entre atacantes y defensores.

- **Ataques biológicos y químicos:** Los recientes sistemas de IA de propósito general han demostrado cierta capacidad para proporcionar instrucciones y orientación para la resolución de problemas con el fin de reproducir armas biológicas y químicas conocidas, y facilitar el diseño de nuevos compuestos tóxicos. En nuevos experimentos que probaron la capacidad de generar planes para producir armas biológicas, un sistema de IA de propósito general obtuvo en ocasiones mejores resultados que expertos humanos con acceso a Internet. En respuesta a esto, una empresa de IA aumentó la evaluación del riesgo biológico de su mejor modelo de "bajo" a "medio". Sin embargo, los intentos reales de desarrollar tales armas aún requieren recursos y conocimientos adicionales. Se vuelve complejo realizar una evaluación exhaustiva del riesgo biológico y químico porque gran parte de la investigación pertinente está clasificada.

Desde la publicación del Informe provisional (mayo de 2024), la capacidad de la IA de propósito general aumentó en dominios que son relevantes para el uso malicioso. Por ejemplo, los investigadores han construido recientemente sistemas de IA de propósito general que fueron capaces de encontrar y explotar algunas vulnerabilidades de ciberseguridad por sí solos y, con asistencia humana, pudieron descubrir una vulnerabilidad previamente desconocida en un software ampliamente utilizado. Las capacidades de la IA de propósito general también han mejorado en relación con el razonamiento y la integración de diferentes tipos de datos, lo que ayuda en la investigación sobre patógenos o en otros campos de doble uso.

Riesgos por fallos: **La IA de propósito general también puede causar daños no deseados.** Incluso cuando los usuarios no tienen intención de causar daño, pueden surgir riesgos graves debido al mal funcionamiento de la IA de propósito general. Entre estos malos funcionamientos se incluyen los siguientes:

- **Problemas de fiabilidad:** La IA de propósito general actual puede ser poco confiable, lo que puede provocar daños. Por ejemplo, si los usuarios consultan un sistema de IA de propósito general para obtener asesoramiento médico o legal, el sistema podría generar una respuesta con información falsa. Estos usuarios no suelen ser conscientes de las limitaciones de un producto de IA, por ejemplo, debido a una limitada alfabetización en IA, publicidad engañosa o errores de comunicación. Existen numerosos casos de daños derivados de problemas de confiabilidad, pero la evidencia sobre la magnitud de estos tipos de problema sigue siendo limitada.
- **Sesgos:** IA de propósito general pueden amplificar los sesgos sociales y políticos, lo que causa daños concretos. Suelen mostrar sesgos en relación con la raza, el género, la cultura, la edad, las discapacidades, la opinión política u otros aspectos de la identidad humana. Esto puede generar resultados discriminatorios, incluida la asignación desigual de recursos, el refuerzo de estereotipos y el rechazo sistemático de grupos o puntos de vista subrepresentados. Los enfoques técnicos para mitigar el sesgo y la discriminación en los sistemas de IA de propósito general están avanzando pero enfrentan una disyuntiva entre la

mitigación del sesgo y los objetivos competitivos, como la precisión y la privacidad, además de otros desafíos.

- **Pérdida de control:** Los escenarios de "pérdida de control" son escenarios futuros hipotéticos en los que uno o más sistemas de IA de propósito general comienzan a operar fuera del control de cualquier persona, sin un camino claro para recuperar el control. Existe un amplio consenso en cuanto a que la IA de propósito general actualmente no plantea este riesgo por carecer de las capacidades necesarias. Sin embargo, la opinión de los expertos sobre la probabilidad de pérdida de control en los próximos años varía enormemente: algunos lo consideran inverosímil, otros consideran que es probable que ocurra y otros lo ven como un riesgo de probabilidad moderada que merece atención debido a su alta gravedad potencial. Las investigaciones empíricas y matemáticas en curso están haciendo avanzar este debate gradualmente.

Desde la publicación del Informe provisional (mayo de 2024), nuevas investigaciones han aportado conocimientos novedosos sobre los riesgos de sesgo y pérdida de control. La evidencia de sesgos en los sistemas de IA de propósito general ha aumentado y trabajos de investigación recientes han detectado formas adicionales de sesgos en IA. Los investigadores han observado modestos avances adicionales hacia las capacidades de IA que probablemente sean necesarias para que ocurran los escenarios de pérdida de control que se debaten comúnmente. Se incluyen capacidades para utilizar computadoras de forma autónoma, programar, obtener acceso no autorizado a sistemas digitales e identificar formas de evadir la supervisión humana.

***Riesgos sistémicos:* más allá de los riesgos directamente asociados con las capacidades de modelos individuales, el despliegue generalizado de la IA de propósito general está relacionado con varios riesgos sistémicos más amplios.** Los ejemplos de riesgos sistémicos van desde posibles impactos en el mercado laboral hasta riesgos de privacidad y efectos ambientales:

- **Riesgos en el mercado laboral:** La IA de propósito general (en particular, si continúa avanzando rápidamente) tiene el potencial de automatizar una amplia gama de tareas, lo que podría tener un efecto significativo en el mercado laboral. Esto significa que muchas personas podrían perder sus empleos actuales. Sin embargo, muchos economistas esperan que las potenciales pérdidas de empleos puedan compensarse, en parte o incluso totalmente, con la creación de nuevos empleos y una mayor demanda en sectores no automatizados.
- **Brecha global en I+D de IA:** Actualmente, la investigación y el desarrollo en la IA de propósito general se concentra en unos pocos países occidentales y en China. Esta “brecha de la IA” tiene el potencial de aumentar la dependencia de gran parte del mundo en este pequeño conjunto de países. Algunos expertos también esperan que esto contribuya a la desigualdad global. La brecha tiene muchas causas, incluidas algunas que no son exclusivas de la IA. Sin embargo, esto se debe en gran medida a los diferentes niveles de acceso a los

costosos recursos informáticos necesarios para desarrollar sistemas de IA de propósito general: La mayoría de los países de ingresos bajos y medios (PIBM) tienen considerablemente menos acceso a recursos informáticos que los países de ingresos altos (PIA).

- **Concentración del mercado y puntos únicos de fallo:** Actualmente, un pequeño número de empresas dominan el mercado de la IA de propósito general. Esta concentración del mercado podría hacer que las sociedades sean más vulnerables a varios riesgos sistémicos. Por ejemplo, si las organizaciones de sectores críticos, como finanzas o atención médica, dependen todas de una pequeña cantidad de sistemas de IA de propósito general, entonces, un error o una vulnerabilidad en dicho sistema podría causar fallos y disrupciones simultáneas a gran escala.
- **Riesgos medioambientales:** El creciente uso de los recursos informáticos en el desarrollo y el despliegue de la IA de propósito general ha incrementado rápidamente las cantidades de energía, agua y materia prima consumidas en la construcción y el funcionamiento de la infraestructura informática necesaria. A pesar del progreso en las técnicas que permiten usar recursos informáticos de manera más eficiente, esta tendencia no muestra indicios definitivos de desaceleración. La IA de propósito general también tiene una serie de aplicaciones que pueden beneficiar o perjudicar las iniciativas de sostenibilidad.
- **Riesgos de privacidad:** La IA de propósito general puede causar o contribuir a violaciones de la privacidad del usuario. Por ejemplo, la información confidencial presente en los datos de entrenamiento se puede filtrar involuntariamente cuando un usuario interactúa con el sistema. Además, cuando los usuarios comparten información confidencial con el sistema, esta información también se puede filtrar. Además, la IA de propósito general también puede facilitar violaciones deliberadas de la privacidad, por ejemplo, si actores maliciosos utilizan la IA para inferir información sensible sobre individuos específicos a partir de grandes cantidades de datos. Sin embargo, hasta ahora los investigadores no han encontrado evidencia de violaciones generalizadas de la privacidad asociadas con la IA de propósito general.
- **Infracciones de derechos de autor:** La IA de propósito general aprende de las obras de expresión creativa y también las crea, con lo que desafía los sistemas tradicionales de consentimiento, compensación y control de datos. La recopilación de datos y la generación de contenido pueden involucrar una variedad de leyes de derechos de datos, que varían según la jurisdicción y pueden ser objeto de litigio activo. Dada la incertidumbre jurídica en torno a las prácticas de recopilación de datos, las empresas de IA están compartiendo menos información sobre los datos que utilizan. Esta opacidad dificulta la investigación de la seguridad de la IA por parte de terceros.

Desde la publicación del Informe provisional (mayo de 2024), ha surgido evidencia adicional sobre los impactos de la IA de propósito general en el mercado laboral, mientras que los nuevos avances han alimentado las preocupaciones sobre la privacidad y los derechos de autor. Los nuevos análisis de los datos del mercado laboral sugieren que las personas están

adoptando la IA de propósito general muy rápidamente en comparación con las tecnologías anteriores. El ritmo de adopción por parte de las empresas varía ampliamente según el sector. Además, los avances recientes en torno a las capacidades ayudaron a que la IA de propósito general se implemente cada vez más en contextos sensibles como la atención médica o la supervisión del lugar de trabajo, lo que crea nuevos riesgos para la privacidad. Por último, como las disputas sobre derechos de autor se intensifican y las mitigaciones técnicas de las infracciones de derechos de autor siguen siendo poco confiables, los titulares de los derechos de datos han restringido rápidamente el acceso a sus datos.

***Modelos de peso abierto:* un factor importante al evaluar muchos de los riesgos que un modelo de IA de propósito general podría presentar es cómo se pone a disposición del público.** Los llamados "modelos de peso abierto" son modelos de IA cuyos componentes centrales, llamados "pesos", se comparten públicamente para su descarga. El acceso a los pesos abiertos facilita la investigación y la innovación, incluso en la seguridad de la IA, además de aumentar la transparencia y ayudar a la comunidad de investigación a detectar fallas en los modelos. Sin embargo, los modelos de peso abierto también pueden presentar riesgos, por ejemplo, al facilitar usos malintencionados o erróneos difíciles o imposibles de controlar por el desarrollador del modelo. Una vez que los pesos del modelo estén disponibles para su descarga pública, no hay forma de implementar una reversión generalizada de todas las copias existentes ni de garantizar que todas las copias existentes reciban actualizaciones de seguridad. Desde el Informe provisional (mayo de 2024), ha surgido un consenso de alto nivel: se propone evaluar los riesgos que plantea la mayor apertura de la IA en términos de riesgo “marginal”. Es decir, la medida en que el lanzamiento de un modelo de peso abierto aumentaría o disminuiría un riesgo determinado, en relación con los riesgos planteados por alternativas existentes, como modelos cerrados u otras tecnologías.

Sección 3 – Gestión de riesgos: ¿Qué técnicas existen para gestionar los riesgos de la IA de propósito general?

Existen varios enfoques técnicos que pueden ayudar a gestionar los riesgos, pero en muchos casos los mejores enfoques disponibles aún tienen limitaciones muy significativas y no ofrecen estimaciones cuantitativas de los riesgos ni garantías que sí existen en otros dominios críticos para la seguridad.

La gestión de riesgos (identificar y evaluar los riesgos, y luego mitigarlos y controlarlos) resulta difícil en el contexto de la IA de propósito general. Aunque la gestión de riesgos también ha sido un gran desafío en muchos otros dominios, hay algunas características de la IA de propósito general que parecen crear dificultades distintivas.

Varias características técnicas de la IA de propósito general hacen que la gestión de riesgos en este ámbito sea particularmente difícil. Estas características incluyen, entre otras, lo siguiente:

- **La gama de posibles usos y contextos de uso de los sistemas de IA de propósito general es inusualmente amplia.** Por ejemplo, el mismo sistema se puede utilizar para proporcionar asesoramiento médico, analizar códigos informáticos para detectar vulnerabilidades y generar fotografías. Esto aumenta la dificultad de anticipar exhaustivamente los casos de uso relevantes, identificar riesgos o probar cómo se comportarán los sistemas en circunstancias relevantes del mundo real.
- **Los desarrolladores aún entienden poco sobre cómo funcionan sus modelos de IA de propósito general.** Esta falta de comprensión hace que sea más difícil predecir problemas de comportamiento, así como explicar y resolver problemas conocidos una vez que se observan. La comprensión sigue siendo difícil principalmente porque los modelos de IA de propósito general no están programados en un sentido tradicional. En lugar de ello, se los entrena: Los desarrolladores de IA configuran un proceso de entrenamiento que involucra un gran volumen de datos, y el resultado de ese proceso de entrenamiento es el modelo de IA de propósito general. El funcionamiento interno de estos modelos es, en gran medida, incomprensible, incluso para sus desarrolladores. Las técnicas de explicación de los modelos y de la "capacidad de interpretación" pueden mejorar la comprensión de los investigadores y desarrolladores sobre cómo funcionan los modelos de IA de propósito general, pero, a pesar de los avances recientes, esta investigación aún está en sus etapas iniciales.
- **Los agentes de IA (sistemas de IA de propósito general que pueden actuar, planificar y delegar de manera autónoma para lograr objetivos), cada vez más capaces, probablemente presentarán desafíos nuevos y considerables para la gestión de riesgos.** Los agentes de IA normalmente trabajan para alcanzar objetivos de forma autónoma mediante el uso de software general, como navegadores web y herramientas de programación. Actualmente, la mayoría aún no son lo suficientemente confiables para un uso generalizado, pero las empresas se están esforzando para construir agentes de IA más capaces y confiables, y han logrado avances en los últimos meses. Es probable que los agentes de IA se vuelvan cada vez más útiles, pero también pueden exacerbar varios de los riesgos analizados en este informe e introducir dificultades adicionales para la gestión de riesgos. Algunos ejemplos de estos nuevos desafíos potenciales incluyen la posibilidad de que los usuarios no siempre sepan qué están haciendo sus propios agentes de IA, la posibilidad de que los agentes de IA operen fuera del control de cualquier persona, la posibilidad de que atacantes puedan tomar el control de otros agentes y la posibilidad de que las interacciones entre diferentes sistemas de IA creen nuevos riesgos complejos. Los enfoques para gestionar los riesgos asociados con los agentes apenas se están empezando a desarrollar.

Además de los factores técnicos, varios factores económicos, políticos y sociales hacen que la gestión de riesgos en el campo de la IA de propósito general sea particularmente difícil.

Consulte la actualización del presidente sobre los últimos avances en IA después de la redacción de este informe.

- **El ritmo del avance de la IA de propósito general crea un "dilema de evidencia" para los tomadores de decisión.** El rápido avance de las capacidades implica que algunos riesgos surjan a pasos agigantados; por ejemplo, el riesgo de hacer trampa en el ámbito académico utilizando IA de propósito general pasó de ser insignificante a generalizado en el plazo de un año. Cuanto más rápido surge un riesgo, más difícil es gestionarlo de forma reactiva y más valiosa resulta la preparación. Sin embargo, mientras la evidencia de un riesgo siga siendo incompleta, los tomadores de decisión tampoco pueden saber con certeza si el riesgo surgirá o incluso si ya ha surgido. Esto crea una disyuntiva: La implementación de medidas preventivas o de mitigación temprana puede resultar innecesaria, pero esperar la evidencia concluyente podría dejar a la sociedad vulnerable ante riesgos que surgen rápidamente. Las empresas y los gobiernos están desarrollando sistemas de alerta temprana y marcos de gestión de riesgos que pueden reducir este dilema. Algunos de ellos activan medidas de mitigación específicas cuando surge nueva evidencia de riesgos, mientras que otros requieren que los desarrolladores proporcionen evidencia de seguridad antes de lanzar un nuevo modelo.
- **Existe una brecha de información entre lo que saben las empresas de IA sobre sus sistemas de IA y lo que saben los gobiernos y los investigadores no industriales.** Las empresas a menudo comparten información limitada sobre sus sistemas de IA de propósito general, especialmente en el período anterior a su lanzamiento general. Las empresas nombran muchas preocupaciones comerciales y de seguridad como razones para limitar el intercambio de información. Sin embargo, esta brecha de información también dificulta la participación efectiva de otros actores en la gestión de riesgos, especialmente en el caso de los riesgos emergentes.
- **Tanto las empresas de IA como los gobiernos suelen enfrentar una fuerte presión competitiva, lo que puede hacer que no prioricen la gestión de riesgos.** En algunas circunstancias, la presión competitiva puede incentivar a las empresas a invertir menos tiempo u otros recursos en la gestión de riesgos de lo que harían en otras circunstancias. De manera similar, los gobiernos pueden invertir menos en políticas de apoyo a la gestión de riesgos en los casos en que perciben una disyuntiva entre la competencia internacional y la reducción de riesgos.

No obstante, existen diversas técnicas y marcos de trabajo para gestionar los riesgos derivados de la IA de propósito general que las empresas pueden emplear y los reguladores pueden exigir. Incluyen métodos para identificar y evaluar riesgos, así como métodos para mitigarlos y controlarlos.

- **Evaluar los sistemas de IA de propósito general en cuanto a riesgos es una parte integral de la gestión de riesgos, pero las evaluaciones de riesgos existentes son sumamente limitadas.** Las evaluaciones existentes de los riesgos de la IA de propósito general se basan principalmente en "controles puntuales", es decir, en probar el comportamiento de la IA de propósito general en un conjunto de situaciones específicas. Esto puede ayudar a detectar peligros potenciales antes de desplegar un modelo. Sin embargo, las pruebas existentes a

menudo pasan por alto los peligros y sobreestiman o subestiman las capacidades y los riesgos de la IA de propósito general, ya que las condiciones de prueba difieren del mundo real.

- **Para que la identificación y evaluación de riesgos sea efectiva, los evaluadores necesitan una considerable experiencia, recursos y acceso suficiente a información relevante.** Una evaluación de riesgos rigurosa en el contexto de la IA de propósito general requiere la combinación de múltiples enfoques de evaluación. Estos van desde análisis técnicos de los modelos y sistemas hasta evaluaciones de posibles riesgos derivados de determinados patrones de uso. Los evaluadores necesitan una experiencia considerable para realizar dichas evaluaciones correctamente. Para realizar evaluaciones de riesgos exhaustivas, también suelen necesitar más tiempo, un acceso más directo a los modelos y sus datos de entrenamiento y más información sobre las metodologías técnicas utilizadas que la que suelen proporcionar las empresas que desarrollan IA de propósito general.
- **Ha habido avances en el entrenamiento de modelos de IA de propósito general para que funcionen de manera más segura, pero ningún método actual puede prevenir de manera confiable incluso salidas abiertamente inseguras.** Por ejemplo, una técnica llamada "entrenamiento adversario" implica exponer deliberadamente modelos de IA a ejemplos diseñados para hacer que fallen o funcionen mal durante el entrenamiento, con el objetivo de generar resistencia a tales casos. Sin embargo, los adversarios aún pueden encontrar nuevas formas ("ataques") de eludir estas medidas de seguridad con un nivel de esfuerzo bajo o moderado. Además, evidencia reciente sugiere que los métodos de entrenamiento actuales, que dependen en gran medida de una retroalimentación humana imperfecta, pueden ayudar inadvertidamente a los modelos a engañar a los humanos en preguntas complejas, al hacer que los errores sean más difíciles de detectar. Mejorar la cantidad y la calidad de esta retroalimentación es una forma de avanzar, aunque las técnicas de entrenamiento incipientes que utilizan IA para detectar comportamientos engañosos también muestran resultados prometedores.
- **La monitorización, que consiste en identificar riesgos y evaluar el rendimiento una vez que un modelo ya está en uso, y diversas intervenciones para prevenir acciones dañinas, pueden mejorar la seguridad de una IA de propósito general después de que se haya desplegado a los usuarios.** Las herramientas actuales pueden detectar contenido generado por IA, realizar un seguimiento del rendimiento del sistema e identificar entradas/salidas potencialmente perjudiciales, aunque usuarios moderadamente capacitados suelen eludir estas protecciones. Varias capas de defensa que combinan capacidades de intervención y control técnico con supervisión humana mejoran la seguridad, pero pueden ocasionar costos y demoras. En el futuro, los mecanismos habilitados por hardware podrían ayudar a los clientes y reguladores a controlar los sistemas de IA de propósito general de manera más efectiva durante el despliegue y, potencialmente, ayudar a verificar acuerdos de forma transfronteriza, pero aún no existen mecanismos confiables de este tipo.
- **Existen múltiples métodos a lo largo del ciclo de vida de la IA para salvaguardar la privacidad.** Estos incluyen la eliminación de información confidencial de los datos de entrenamiento, enfoques de entrenamiento de modelos que controlan cuánta información

se aprende de los datos (como los enfoques de "privacidad diferencial") y técnicas para usar IA con datos sensibles que dificultan la recuperación de los datos (como el "cálculo confidencial" y otras tecnologías que mejoran la privacidad). Muchos métodos de mejora de la privacidad de otros campos de investigación aún no son aplicables a los sistemas de IA de propósito general debido a los requisitos informáticos de estos sistemas. En los últimos meses, los métodos de protección de la privacidad se han ampliado para abordar el uso creciente de la IA en dominios sensibles, incluidos los asistentes de teléfonos inteligentes, los agentes de IA, los asistentes de voz que siempre escuchan y el uso en la atención médica o la práctica legal.

Desde la publicación del Informe provisional (mayo de 2024), los investigadores han logrado avances adicionales para poder explicar por qué un modelo de IA de propósito general produce un resultado determinado. Poder explicar las decisiones de la IA podría ayudar a gestionar los riesgos derivados del mal funcionamiento que van desde sesgos e inexactitudes fácticas hasta pérdida de control. Además, recientemente se llevaron a cabo iniciativas para estandarizar los enfoques de evaluación y mitigación en todo el mundo.

Conclusión: Existe una amplia gama de trayectorias para el futuro de la IA de propósito general y mucho dependerá de cómo actúen las sociedades y los gobiernos

El futuro de la IA de propósito general es incierto, con una amplia gama de trayectorias posibles hasta en el futuro cercano, incluidos resultados muy positivos y muy negativos. Pero nada sobre el futuro de la IA de propósito general es inevitable. La forma en que se desarrolla la IA de propósito general y quién la desarrolla, qué problemas se diseñan para resolverlos, si las sociedades podrán aprovechar el pleno potencial económico de la IA de propósito general, quién se beneficia de ella, los tipos de riesgos a los que nos exponemos y cuánto invertimos en investigación para gestionar los riesgos: estas y muchas otras preguntas dependen de las decisiones que las sociedades y los gobiernos tomen hoy y en el futuro para dar forma al desarrollo de la IA de propósito general.

Para ayudar a facilitar un debate constructivo sobre estas decisiones, este informe proporciona una descripción general del estado actual de la investigación científica y el debate sobre la gestión de los riesgos de la IA de propósito general. Las implicaciones son grandes. Esperamos continuar con este esfuerzo.