Informe preliminar del Panel Científico Internacional Independiente sobre Inteligencia Artificial

Evaluación con base empírica de las oportunidades, los riesgos y las repercusiones de la inteligencia artificial

Julio de 2026

United Nations
Independent International Scientific Panel on Artificial Intelligence

Informe preliminar del Panel Científico Internacional Independiente sobre Inteligencia Artificial: evaluación con base empírica de las oportunidades, los riesgos y las repercusiones de la inteligencia artificial

Members of the Independent International Scientific Panel on Artificial Intelligence

Girmaw Abebe Tadesse, Ethiopia

Tuka Alhanai, United Arab Emirates

Joëlle Barral, France

Yoshua Bengio, Canada *(Co-Chair)*

Tegawendé Bissyandé, Burkina Faso

Loreto Bravo, Chile

Mark Coeckelbergh, Belgium

Carlos Coello Coello, Mexico

Melahat Bilge Demirköz, Türkiye

Adji Bousso Dieng, Senegal

Awa Bousso Dramé, Cabo Verde

Mennatallah El-Assady, Egypt

Hoda Heidari, Islamic Republic of Iran

Juho Kim, Republic of Korea

Anna Korhonen, Finland

Aleksandra Korolova, Latvia

Vipin Kumar, United States

Sonia Livingstone, United Kingdom

Qinghua Lu, Australia

Teresa Ludermir, Brazil

Vukosi Marivate, South Africa

Bilal Mateen, Pakistan

Yutaka Matsuo, Japan

Joyce Nakatumba Nabende, Uganda

Andrei Neznamov, Russian Federation

Maximilian Nickel, Germany

Rita Orji, Nigeria

Román Orús, Spain

Alvitta Ottley, Saint Kitts and Nevis

Martha Palmer, United States

Johanna Pirker, Austria

Balaraman Ravindran, India

Maria Ressa, Philippines *(Co-Chair)*

Lior Rokach, Israel

Piotr Sankowski, Poland

Silvio Savarese, Italy

Bernhard Schölkopf, Germany

Haitao Song, China

Leslie Teo, Singapore

Jian Wang, China

Acerca de este informe

El presente informe ofrece una evaluación científica independiente preliminar de las capacidades y de las oportunidades y los riesgos emergentes de la inteligencia artificial (IA), y proporciona una base empírica común para ayudar a los Estados Miembros a orientarse en un ámbito tecnológico en rápida evolución. El informe ha sido elaborado por el Panel Científico Internacional Independiente sobre Inteligencia Artificial, órgano que creó la Asamblea General en 2025 mediante su resolución 79/325. El Panel, que es el primer organismo científico mundial dedicado a la IA, opera bajo un mandato estrictamente científico y apolítico con el fin de documentar el consenso y las discrepancias científicas a nivel internacional, procurando mantener la pertinencia de su labor para las políticas, pero sin prescribirlas. El informe, que es el primero de este tipo, irá actualizándose a lo largo del año mediante resúmenes temáticos que tratarán los acontecimientos a medida que se produzcan. Se recogen en él los mejores datos disponibles en el momento de la publicación, en un campo que evoluciona tan rápidamente que cualquier resumen requiere un compromiso de someterlo a revisión.

Descargo de responsabilidad:

El presente informe es obra del Panel Científico Internacional Independiente sobre IA, que es responsable del contenido y la publicación del informe. Los miembros del Panel actúan a título personal y no como representantes de ningún Gobierno ni de ninguna otra autoridad u organización.

El Panel ha actuado de forma totalmente discrecional en lo que respecta a la inclusión del contenido del presente informe. El informe y su contenido responden a un amplio consenso entre los miembros; no se espera que ningún miembro esté de acuerdo con todos y cada uno de los puntos que figuran en el documento. Los miembros manifiestan su acuerdo en el sentido amplio, aunque no unilateral, con las constataciones.

El informe no refleja las opiniones de las Naciones Unidas, y nada de lo contenido en él podrá interpretarse como una expresión de las opiniones o posiciones oficiales de ningún Gobierno ni de ninguna otra autoridad u organización a la que pueda estar vinculado un miembro del Panel.

Las denominaciones, incluidos los nombres geográficos, empleadas en la presente publicación y la forma en que aparecen presentados los datos, como las citas, los mapas y la bibliografía, no entrañan, de parte de las Naciones Unidas, juicio alguno sobre el nombre o la condición jurídica de ninguno de los países, territorios, ciudades o zonas citados o de sus autoridades, ni respecto del trazado de sus fronteras o límites, y no cuentan necesariamente con la aprobación o aceptación oficial de las Naciones Unidas. La presentación de información dimanante de medidas y decisiones adoptadas por los Estados no supone que las Naciones Unidas aprueben, acepten o reconozcan oficialmente esas medidas y decisiones, y tal información se incluye sin perjuicio de la postura de los distintos Estados Miembros de las Naciones Unidas. La información que figura en el presente informe relativa a determinadas sociedades mercantiles o nombres comerciales de ciertos productos no implica ninguna aprobación ni recomendación oficial por parte de las Naciones Unidas (ni siquiera en virtud de la omisión de referencias a otros productos análogos). La presente publicación no afecta a la posición de ningún Estado Miembro de las Naciones Unidas con respecto a ningún acuerdo multilateral internacional.

Índice

Resumen

El presente documento tiene por objeto ofrecer un análisis equilibrado de los riesgos y las oportunidades de la inteligencia artificial (IA). "Equilibrado" significa que existe el compromiso de evaluar los datos empíricos sin un sesgo indebido hacia el optimismo o el pesimismo. Las ventajas potenciales de la IA son enormes. Si se despliega y aplica de forma reflexiva, la IA puede contribuir al avance hacia la consecución de los Objetivos de Desarrollo Sostenible, impulsar el progreso de las ciencias de la salud y mejorar el acceso a la educación. Al mismo tiempo, el rápido ritmo del desarrollo tecnológico y la amplitud de las posibles aplicaciones plantean importantes desafíos a los responsables de formular políticas. El despliegue rápido y descontrolado de esta tecnología en gran escala también plantea riesgos considerables, como los efectos negativos para la salud mental de los usuarios, su posible uso como herramienta destructiva, las repercusiones en los sistemas sociales, económicos y ambientales y las dificultades asociadas al control de la tecnología. El presente informe no pretende abarcar todas las oportunidades y riesgos posibles, sino que se centra en algunos de los más acuciantes.

Capacidades y adopción

En los últimos años se han producido avances rápidos y, en algunos ámbitos, cada vez más acelerados en una amplia gama de capacidades de la IA. Las importantes inversiones en potencia computacional, nuevas metodologías de IA y datos de entrenamiento especializados han dado lugar a mejoras sostenidas en una amplia gama de capacidades de la IA. Son ejemplo de ello la conversación fluida, la generación de código funcional, el razonamiento de nivel experto en matemáticas y ciencias, el análisis de datos en gran escala y la generación de contenido de imagen, audio y video. Siguen existiendo limitaciones, como la fiabilidad, la capacidad de obtener un buen rendimiento en diferentes idiomas y culturas, la interacción con sistemas físicos, la ejecución de proyectos complejos o en varias etapas y la generación de resultados objetivos; sin embargo, en general, desde hace ya varios años el progreso técnico en muchos ámbitos importantes ha procedido con rapidez, superando las expectativas habituales en materia de adelantos tecnológicos.

Estos avances han dado lugar a aplicaciones útiles en los ámbitos de la ciencia, la salud, la agricultura, la accesibilidad, el trabajo intelectual y la tecnología de la información, incluido el propio desarrollo de la IA. Por ejemplo, en el ámbito científico, AlphaFold ha predicho las estructuras de más de 200 millones de proteínas, resultados que ahora utilizan más de 3 millones de investigadores, y ha acelerado el diseño de fármacos, el desarrollo de vacunas y la investigación sobre la resistencia a los antibióticos. En radiología también se ha utilizado la IA para detectar el cáncer de mama en una fase más temprana, mientras que los profesionales sanitarios de primera línea en entornos con escasos recursos utilizan herramientas de inteligencia artificial adaptadas a los idiomas locales para prestar servicios de atención de la salud de mayor calidad.

La adopción de la IA se ha acelerado de forma generalizada, aunque desigual, en los distintos países y sectores. A nivel mundial, más de 1.000 millones de personas utilizan actualmente la IA conversacional cada semana. Sin embargo, el acceso a la IA y su uso varían considerablemente en todo el mundo, y su adopción en el Sur Global va muy por detrás de la del Norte Global. Además, existen diferencias significativas entre las economías avanzadas con respecto a la infraestructura de cómputo y los modelos. Esta disparidad refleja las desigualdades existentes y puede incluso agravarlas. El propio desarrollo de la IA está aún más concentrado: según estimaciones recientes, los Estados Unidos de América suponen el 75 % de la potencia computacional de las 500 supercomputadoras de IA más potentes del mundo, mientras que China supone el 15 %. Las empresas de los Estados Unidos y China también desarrollan

casi todos los modelos de propósito general más importantes, y un pequeño número de países controla los insumos fundamentales para la cadena de suministro de los chips informáticos destinados a la IA.

Aunque la transición hacia los agentes de IA ya está en marcha, es probable que en el futuro su adopción y sus repercusiones económicas dependan de las mejoras continuas en su capacidad para realizar tareas intelectuales con poca o ninguna supervisión humana. Un agente de IA es un sistema informático capaz de planificar y actuar de forma autónoma para alcanzar objetivos, utilizando las herramientas de que dispone. Estos sistemas han mejorado rápidamente en los últimos años: según un estudio, la extensión de determinadas tareas de programación informática que pueden realizar los sistemas más avanzados ha ido duplicándose cada cuatro a siete meses. Si este ritmo se mantiene, los agentes de IA pronto podrán realizar tareas que actualmente llevan días o semanas a los programadores humanos. Dado que pueden funcionar con poca supervisión y a gran velocidad, los agentes de IA pueden reportar importantes beneficios económicos y científicos. Por ejemplo, los sistemas de IA agéntica en laboratorios químicos autónomos han demostrado que la velocidad del descubrimiento de materiales se ha multiplicado por más de diez. La selección de publicaciones con ayuda de la IA podría haber reducido la carga de trabajo en aproximadamente un 60 % en algunos entornos de investigación. Por lo tanto, los agentes de IA tienen importantes repercusiones en todos los sectores. Al mismo tiempo, su despliegue plantea cuestiones urgentes en relación con los mercados laborales, la ciberseguridad, el ecosistema de la información y la gobernanza y controlabilidad de los futuros sistemas de IA.

Comprender y gestionar los riesgos

El desarrollo de la IA conlleva riesgos, como las posibles repercusiones negativas en los derechos humanos, los sistemas sociales y el medio ambiente. Por ejemplo, el material de abuso sexual de niños generado por IA y la violencia sexual facilitada por las ultrafalsificaciones han pasado a circular con mayor frecuencia por Internet, lo que perjudica de manera desproporcionada a mujeres y niños. El comportamiento adulador de la IA, en que las respuestas de la IA refuerzan las creencias preexistentes de los usuarios independientemente de su veracidad, se ha relacionado con varios incidentes graves de salud mental, incluidas muertes documentadas. La IA facilita la producción y la difusión en gran escala de contenidos persuasivos, incluidos los concebidos para causar engaño, lo que contribuye a una erosión gradual de la integridad de la información que puede debilitar la realidad compartida necesaria para la confianza pública, la cohesión social y el debate democrático. Se ha constatado que delincuentes y agentes dañinos utilizan sistemas de IA para facilitar la realización de ciberataques. Muchos de estos perjuicios recaen de manera desproporcionada en poblaciones que ya se encuentran en situación de desventaja.

De cara al futuro, el desfase entre unas capacidades en rápida evolución y unos métodos de gestión de riesgos que puedan ser eficaces podría tener consecuencias catastróficas. Por ejemplo, las habilidades técnicas avanzadas pueden permitir que actores privados sin experiencia utilicen la IA con fines maliciosos en una amplia gama de ámbitos, como el fraude, la ingeniería social, la ciberseguridad, la desinformación, la biotecnología y la manipulación financiera. No existen métodos fiables para mantener el control sobre sistemas de IA altamente autónomos. No hay garantías científicas de que los agentes de IA no vayan a incumplir las instrucciones, y cada vez se documentan más casos en que ya las incumplen. En entornos de laboratorio, se ha demostrado que los sistemas de IA incumplen sus instrucciones de seguridad para evitar que se les apague. Este tipo de comportamiento puede plantear dificultades para los métodos de evaluación y supervisión, ya que aumenta la capacidad de los principales sistemas de IA para reconocer los entornos de pruebas y generar resultados de evaluación

engañosos que favorecerían su funcionamiento continuado. Además, pueden surgir nuevos riesgos a raíz de las interacciones entre múltiples agentes.

Los riesgos de la IA se distribuyen de forma desigual entre las poblaciones y los países, mientras que el desarrollo de la IA y la riqueza que genera están muy concentrados. La concentración de las capacidades de IA en un reducido número de empresas y países podría facilitar la injerencia de regímenes autoritarios y socavar la rendición de cuentas democrática.

Establecer la gobernanza de la inteligencia artificial para aprovechar sus ventajas y mitigar sus riesgos

Para aprovechar al máximo las ventajas de la IA y minimizar sus riesgos es necesario contar con una buena gobernanza. Los avances económicos y laborales y su distribución equitativa no se producen de forma automática: si se realizan inversiones complementarias en habilidades, flujos de trabajo, infraestructura e instituciones del mercado de trabajo, la tecnología puede crear nuevos empleos que aún no existen, como demuestra el hecho de que más del 60 % de los empleos en 2018 eran nuevos en comparación con los de 1945. Si faltan estas inversiones, se corre el riesgo de que la IA agrave las desigualdades, desplace a los trabajadores y desvíe la riqueza del trabajo hacia el capital, en lugar de crear empleos sostenibles y de calidad, que ofrezcan una remuneración justa, autonomía a los trabajadores y un camino seguro hacia la dignidad social. La IA puede ampliar enormemente las capacidades humanas al ofrecer educación personalizada, herramientas de salud mental accesibles y tecnologías de apoyo mejoradas, pero para aprovechar estas oportunidades de forma segura se necesita realizar inversiones específicas y disponer de políticas con que se fomente un acceso equitativo y se recompense la innovación, al tiempo que se evita la explotación de las poblaciones vulnerables, en particular los niños, y se previene la pérdida de conocimientos especializados, la dependencia psicológica o la aniquilación cultural y lingüística.

Los responsables de formular políticas que pretenden dar forma a esta gobernanza se encuentran ante un dilema empírico: necesitan una base empírica para adoptar decisiones de gobernanza bien fundamentadas y de importancia trascendental, pero para cuando esa base esté disponible podría ser demasiado tarde para adoptarlas, ya que las pruebas empíricas van por detrás del ritmo de desarrollo de la IA. En diversas jurisdicciones ya están utilizándose decenas de instrumentos de gobernanza distintos que tienen por objeto integrar la ética y los derechos humanos en los sistemas de IA, pero estos instrumentos se encuentran fragmentados, se concentran en unas pocas empresas y su eficacia en la práctica rara vez se evalúa. Los propios métodos de evaluación están poco desarrollados, y las instituciones necesarias para realizar evaluaciones independientes de la capacidad y los riesgos siguen en estado incipiente.

La capacidad para actuar en función de las pruebas empíricas disponibles sobre los riesgos y las repercusiones de la IA está distribuida de forma desigual. La mayoría de los países, incluidas muchas economías avanzadas, carecen de los conocimientos técnicos necesarios para evaluar los modelos “de frontera” más potentes o participar de manera efectiva en su gobernanza. La infraestructura de cómputo, los conocimientos especializados en materia de evaluación y los datos (por ejemplo, para abarcar diferentes idiomas) se concentran allí donde se desarrolla la IA, lo que hace que la mayoría de los Estados Miembros dependan de sistemas que no pueden crear, inspeccionar, auditar ni adaptar plenamente al contexto local. El mero acceso a las herramientas de IA no garantiza que todos se beneficien por igual; las inversiones complementarias en datos, habilidades, flujos de trabajo e instituciones que permitan pasar del acceso a un despliegue útil, costoeficaz y seguro son necesarias, pero su distribución es dispar.

Existen medidas concretas para subsanar las carencias mencionadas, pero cada una de ellas requiere una inversión sostenida en la capacidad de los

Estados Miembros para configurar, evaluar y desplegar la IA. Este informe preliminar forma parte, en sí mismo, de la contribución del Panel, una base empírica compartida para los Estados Miembros que deben adoptar decisiones cada vez más urgentes. A medida que vaya profundizando en su comprensión gracias a la interacción continua con los Estados Miembros y la comunidad científica en general, el Panel también perfeccionará su análisis, que irá más allá de las carencias encontradas para deslindar las trayectorias, las tensiones y las oportunidades que definirán el futuro de la IA.

1. ¿Por qué en este momento?

La realidad actual

Nos encontramos en un punto de inflexión. La inteligencia artificial no es tan solo una de tantas tecnologías emergentes; es la primera en reducir el tiempo de adopción de décadas a meses, industrializar el trabajo cognitivo en gran escala y concentrar la capacidad de transformación en manos de unos pocos actores mundiales. Este informe proporciona a los responsables de formular políticas de todas las regiones la base empírica común que se necesita para dar respuesta a esta cuestión.

¿Qué es la inteligencia artificial?

La inteligencia artificial (IA) es una tecnología transformadora, pero también es cierto que el objetivo al que se encamina no se mantiene invariable. El concepto ha ido evolucionando con el tiempo, pasando de la IA simbólica al aprendizaje automático, la IA generativa y la IA agéntica, y en ocasiones incluso a la inteligencia artificial general o la superinteligencia.

Los sistemas de IA son sistemas informáticos que, por decirlo de algún modo, perciben, aprenden y actúan. A partir de los datos que se les dan, deducen cómo generar resultados tales como predicciones, contenidos, recomendaciones, acciones o decisiones, con distintos grados de autonomía y capacidad de adaptación. El denominador común de la IA actual, más que una arquitectura concreta, es el hecho de que los sistemas modernos aprenden de la experiencia representada por los datos. Esa experiencia adopta diversas formas: el aprendizaje a partir de huellas culturales humanas (textos, imágenes, código) constituye la base del "preentrenamiento" de los modelos fundacionales actuales, que son sistemas de gran tamaño y ampliamente entrenados que sustentan una amplia gama de aplicaciones de IA; el aprendizaje mediante la interacción con el mundo (digital y físico) es la base del aprendizaje de refuerzo y la robótica; y el aprendizaje a partir de simulaciones permite a los agentes adquirir experiencia en entornos virtuales.

Modelos fundacionales e IA específica para tareas concretas. El debate público suele centrarse en los modelos fundacionales y la IA de propósito general, que se caracterizan por su capacidad para realizar —o adaptarse para realizar— una amplia variedad de tareas. Estos sistemas se están desplegando actualmente en gran escala y constituyen el tema principal del presente informe. Sin embargo, muchos sistemas de IA de tarea específica (IA estrecha) están diseñados para realizar tareas concretas en ámbitos restringidos. Esta distinción es importante. La IA de tarea específica ofrece ventajas cuantificables cuando la tarea está bien definida, se dispone de los datos necesarios y las instituciones pueden desplegarla de forma planificada. Los sistemas de propósito general ofrecen flexibilidad, pero plantean diversas cuestiones en materia de gobernanza.

¿En qué se diferencia la inteligencia artificial de otras tecnologías emergentes?

Un ritmo de adopción sin precedentes. Los sistemas de IA se consideran cada vez más una tecnología de propósito general, tan transformadora en la amplitud de sus aplicaciones como la máquina de vapor, la electricidad e Internet [1, 2]. Sin embargo, su caso es diferente desde varios puntos de vista importantes. La electricidad tardó décadas en llegar a la mayoría de los hogares; la globalización de Internet mediante la World Wide Web tardó unos 15 años en alcanzar 1.000 millones de usuarios. ChatGPT alcanzó 100 millones de usuarios en dos meses [3]. La formulación de políticas tradicional no ha podido seguir el ritmo [4].

Concentración y homogeneización. La IA moderna, en particular los modelos fundacionales, demuestra unas economías de escala sin precedentes que generan una fuerte presión hacia la centralización de la capacidad, ya que los sistemas más potentes requieren un entrenamiento con recursos computacionales a los que solo tienen acceso unos pocos actores a nivel mundial. Esta concentración de recursos conlleva el riesgo de una homogeneización del conocimiento y la cultura. Por ejemplo, el entrenamiento con conjuntos de datos consolidados puede dar lugar a sistemas que reflejen de forma sistemática los idiomas y las perspectivas dominantes, marginando al resto.

Una repercusión enorme en el trabajo intelectual. Mientras que las primeras oleadas de automatización transformaron el trabajo físico, la IA es la primera en afectar de forma masiva al trabajo cognitivo y creativo, incluyendo la redacción, la programación, el análisis jurídico, el diagnóstico médico, los descubrimientos científicos, la elaboración de previsiones y la generación de imágenes [1, 2, 5–8].

Dificultad de garantizar la veracidad y la corrección de los resultados. Los modelos de IA actuales aprenden a predecir patrones en grandes conjuntos de datos. Los modelos de lenguaje no solo aprenden plantillas almacenadas, sino también transformaciones mediante las cuales puede convertirse una oración en muchas formas conexas, conservando al mismo tiempo la fluidez. Esto permite generar resultados novedosos, en lugar de limitarse a copiar, pero da lugar a una asimetría fundamental: es más fácil producir un texto fluido que un texto basado en hechos [8]. Los grandes modelos de lenguaje pueden presentar alucinaciones como si fueran hechos y los usuarios suelen interpretar la soltura en el manejo del lenguaje como una prueba de veracidad y fidelidad a los hechos. Esta desconexión entre la aparente competencia y la exactitud marca el panorama de riesgos de manera sistemática. En el ámbito de la atención de la salud, la IA de propósito general reduce el tiempo dedicado a la documentación administrativa, una tarea en que se valora la capacidad de la IA para sintetizar y estructurar la información. Sin embargo, según los datos, una de cada cuatro conversaciones con chatbots se refiere a la salud o el bienestar [9], lo que significa que estos mismos sistemas con frecuencia se consultan con fines diagnósticos, donde la exactitud fáctica es fundamental y los errores pueden acarrear graves consecuencias. La tecnología es la misma, pero lo que está en juego no lo es. El que algo sea viable no quiere decir que sea idóneo, por lo que su despliegue sin un seguimiento y una evaluación sistemáticos, especialmente en ámbitos que por el momento carecen de marcos regulatorios, conlleva el riesgo de causar daños difíciles de detectar.

La urgente necesidad de realizar una evaluación independiente y con base empírica de los sistemas de inteligencia artificial

La necesidad de una evaluación científica independiente en este momento viene determinada por unas circunstancias que alteran lo que está en juego desde el punto de vista regulatorio.

En primer lugar, el desarrollo de la IA está avanzando a un ritmo superior al previsto en las evaluaciones previas de los expertos y los ciclos regulatorios. Se sigue avanzando en las capacidades en ámbitos clave, impulsados por nuevas técnicas de entrenamiento y el escalado de cómputo en tiempo de inferencia [10]. Las mediciones empíricas muestran que las capacidades de la IA se están acelerando [11].

En segundo lugar, los principales desarrolladores de modelos de frontera han comenzado a restringir el despliegue de modelos que superan los umbrales de riesgo definidos internamente [12–15]. Sin embargo, estos umbrales siguen siendo establecidos por los propios desarrolladores, sin que exista una evaluación estandarizada ni una verificación externa [16–17].

En tercer lugar, los enfoques de gobernanza de la IA siguen estando fragmentados entre las distintas regiones. Se observa un creciente desorden en la gobernanza global, ya que algunos países han promulgado leyes específicas sobre la IA con

normas fundamentalmente contradictorias y costos de cumplimiento. Las jurisdicciones presentan filosofías de regulación divergentes, carecen de un mecanismo unificado de gestión de riesgos y de criterios de evaluación comparables y no se coordinan entre ellas más que de forma limitada, lo que conlleva el riesgo de un panorama regulatorio fragmentado [18]. Sin embargo, la fragmentación no es una fatalidad, y aún existe la oportunidad de establecer normas comunes en materia de pruebas empíricas y coordinar la supervisión a nivel mundial.

¿Qué hace que el Panel Científico Internacional Independiente sobre Inteligencia Artificial sea único?

Numerosos grupos de expertos vinculados a organizaciones internacionales, consejos científicos nacionales, consorcios industriales y centros de investigación independientes elaboran evaluaciones sobre el desarrollo de la IA. Estas iniciativas son valiosas, pero se ven limitadas por su ámbito geográfico, su enfoque sectorial o su falta de continuidad.

Este Panel ocupa una posición especial por tres motivos. En primer lugar, el Panel parte de la premisa de que las Naciones Unidas constituyen el principal foro mundial sobre riesgos transfronterizos de esta magnitud, como se expone en el informe *Gobernanza de la IA en beneficio de la humanidad* y se refleja en el Pacto Digital Global, que establece la necesidad imperiosa de reconocer que "la rápida evolución y la potencia de las tecnologías emergentes están creando nuevas posibilidades" y "determinar y mitigar los riesgos y garantizar la supervisión humana de la tecnología para promover el desarrollo sostenible y el pleno disfrute de los derechos humanos".

En segundo lugar, el Panel es actualmente el único mecanismo permanente de las Naciones Unidas con el mandato de realizar evaluaciones científicas periódicas sobre el estado, los riesgos y las capacidades de la IA, y está concebido para llevar a cabo una labor continuada e iterativa.

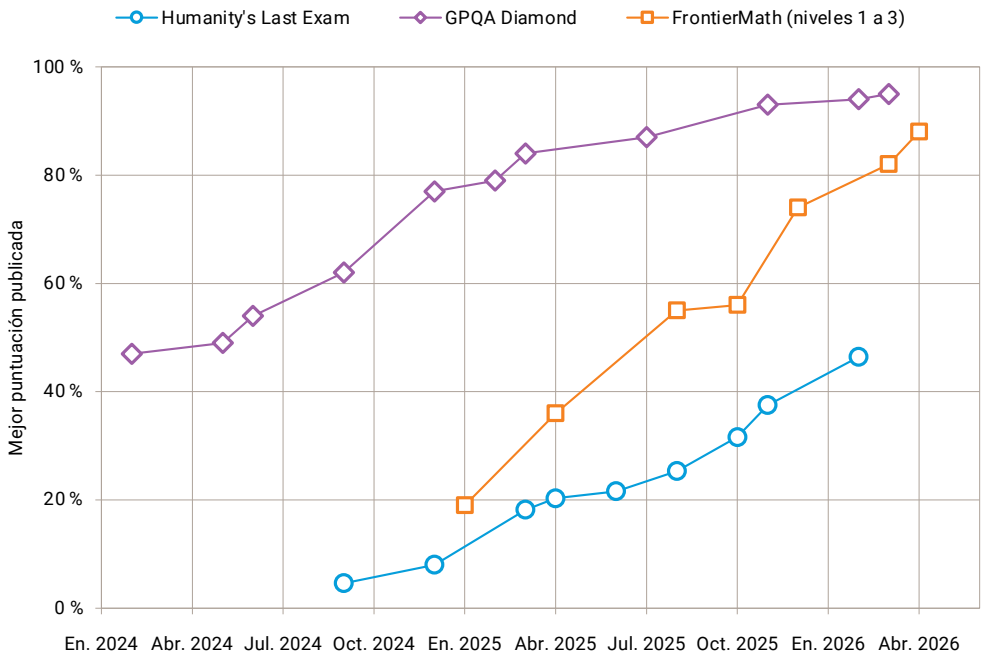
En tercer lugar, el Panel tiene un mandato científico, no político: documentar las pruebas científicas, el consenso y los desacuerdos, así como las lagunas de conocimiento que sigue siendo urgente subsanar. Su objetivo es proporcionar a los Gobiernos y las instituciones la base empírica que necesitan para actuar en los próximos meses y años, manteniendo la pertinencia de su labor para las políticas, pero sin prescribirlas. Este carácter científico debería hacer que sus conclusiones sean comparables entre regiones y resilientes frente a los ciclos políticos.

2. ¿Qué revelan los datos empíricos?

El rendimiento de la IA medido según las principales pruebas de referencia (benchmarks) ha aumentado extraordinariamente en los últimos años (véase la figura I).

Figura I

Pruebas de referencia para la inteligencia artificial, estado actual de la técnica (desde 2024 hasta mayo de 2026)



Fuente: Adaptado de EpochAI, 2026, https://epoch.ai/benchmarks.

Humanity's Last Exam, una prueba de 2.500 preguntas concebida específicamente para que plantee dificultades a los modelos de propósito general, ha visto cómo las puntuaciones máximas pasaban del 8 % al 45 % en 16 meses [19]. En GPQA Diamond, una prueba de razonamiento científico de nivel de doctorado, en la actualidad los mejores modelos responden correctamente a alrededor del 95 % de las preguntas, frente al 36 % registrado en 2023 [20]. La puntuación máxima obtenida en FrontierMath, una prueba que evalúa la capacidad de razonamiento matemático, pasó del 19 % en enero de 2025 al 88 % en 2026 [20]. Varios sistemas de IA obtuvieron una puntuación equivalente a la medalla de oro en la Olimpiada Internacional de Matemática de 2025, un hito que se ha alcanzado mucho antes de lo que muchos expertos habían pronosticado [21].

2.1 Las capacidades de la inteligencia artificial avanzan más rápido que nuestra capacidad para medirlas o gobernarlas

La medición y la evaluación de la IA constituyen la base para valorar las oportunidades, los riesgos y las repercusiones de la IA. Sin embargo, el ritmo sin precedentes del desarrollo de la IA plantea los siguientes desafíos en materia de valoración:

a) **Existe una asimetría de información entre las empresas y la sociedad en lo que respecta a la validación de la seguridad**. Los desarrolladores de la IA de frontera conservan la visibilidad exclusiva sobre los sistemas que han creado. En la actualidad, las metodologías de evaluación de la seguridad son diseñadas en gran medida por las propias empresas que son objeto de evaluación. Aunque existen algunas obligaciones legales de divulgación, los datos de pruebas que reciben los expertos gubernamentales son principalmente los que los desarrolladores deciden compartir. Sin una evaluación externa estandarizada, rigurosa e independiente, semejante a la que existe en los sectores farmacéutico

y aeronáutico, la garantía de seguridad depende en gran medida de la buena voluntad de los desarrolladores [21];

b) **La IA puede memorizar soluciones a pruebas que están disponibles públicamente**. Si el modelo de IA ha memorizado (inadvertidamente) las respuestas correctas a preguntas de las pruebas como parte del proceso de entrenamiento, es posible que su rendimiento en esas pruebas no pueda extrapolarse a preguntas similares. Para evitar la contaminación de los datos, cada vez es más habitual que los conjuntos de datos de evaluación se mantengan confidenciales [22];

c) **Cada vez hay más pruebas que resultan demasiado fáciles para la IA**. Los modelos de IA obtienen puntuaciones casi perfectas en un número cada vez mayor de los exámenes estandarizados, llamados *benchmarks* o pruebas de referencia, que los investigadores utilizan para evaluar las capacidades de los modelos antes de su lanzamiento [23]. Por lo tanto, las pruebas de referencia afectadas ya no permiten distinguir entre un modelo muy competente y otro aún mejor;

d) **Los modelos de IA son capaces de engañar de forma activa**. El engaño se produce cuando un sistema de IA induce a error de forma sistemática a los seres humanos u otros agentes en relación con sus conocimientos, planes o capacidades. Este fenómeno se observa cada vez más en la práctica. El engaño puede alterar las evaluaciones de seguridad y la fiabilidad en el mundo real, y es pertinente en hipótesis de pérdida de control, como lo demuestran los comportamientos de los modelos de IA que mienten y engañan para evitar que se les apague [24];

e) **Es posible que los modelos de IA se den cuenta de cuándo se les está sometiendo a prueba**. Este nuevo desafío se denomina “conciencia de evaluación” [25]. Si a esto le sumamos la capacidad de engaño, esto significa que los modelos de IA podrían reducir temporalmente su rendimiento en las pruebas de evaluación de capacidades peligrosas, obedeciendo instrucciones de seres humanos o por decisión propia [26];

f) **La IA agéntica complica las pruebas**. Los agentes de IA que actúan en nombre de seres humanos pueden utilizar herramientas para llevar a cabo tareas de larga duración sin supervisión humana directa. Las metodologías que estén calibradas para evaluar la capacidad de un agente para actuar de forma independiente, su impacto en el entorno operacional y su comportamiento emergente están poco desarrolladas [27]. Además, cuando interactúan varios agentes adaptativos, surgen nuevos riesgos sistémicos, como la falta de coordinación, los conflictos y la confabulación [28].

Entre las respuestas a estos desafíos se encuentran las siguientes:

a) **Pruebas dinámicas basadas en la ejecución**. Una medición precisa requiere recursos considerables para desarrollar constantemente nuevas pruebas de referencia que tengan el grado de dificultad suficiente para producir sistemas de IA más avanzados. Para que las pruebas de referencia sigan siendo difíciles y conserven utilidad real, las prácticas de evaluación están pasando de entornos estáticos a entornos dinámicos basados en la ejecución [29, 30]. Sin embargo, crear este tipo de entornos resulta más costoso que elaborar un cuestionario de conocimientos, y son pocos los actores que han invertido lo suficiente en la medición de la IA como para desarrollarlos;

b) **Interpretabilidad**. Los métodos de interpretabilidad, cuyo objetivo es comprender qué ocurre en el interior de los modelos de IA, están cobrando cada vez más importancia para la detección de comportamientos peligrosos ocultos. Un método destacado es el de la “cadena de pensamiento”, mediante el cual el modelo expone los pasos de su razonamiento antes de dar una respuesta. El método resulta prometedor, pero es fundamental asegurar que este razonamiento sea fiel y legible para los seres humanos [31]. Otro enfoque consiste en un

clasificador entrenado a partir de las activaciones internas del modelo, capaz de predecir la respuesta a una pregunta como: "¿Es sincero este modelo?" [32]. Sin embargo, un clasificador de este tipo requiere acceso a las activaciones de los pesos del modelo, y solo puede evaluarse de forma independiente en el caso de modelos de IA cerrados si las organizaciones de evaluación de confianza obtienen un acceso más profundo [33];

c) **Medición continua**. Este método supone seguir el rastro del modo en que se comporta un sistema tras su lanzamiento, con usuarios reales, tareas reales y entornos reales. Este seguimiento posterior a la puesta en el mercado puede incluir patrones de uso de la IA anonimizados y agregados que hayan facilitado los desarrolladores de IA [34], notificaciones de incidentes y resultados comunicados por los usuarios. Hasta la fecha, no existe una norma común para el análisis de los patrones de uso por parte de los productores de IA que garantice la protección de la privacidad. Además, el conocimiento del ecosistema es menor en el caso de los modelos abiertos que pueden descargarse y utilizarse sin que los creadores de la IA tengan conocimiento alguno de ello. Los proveedores de sistemas de IA de alto riesgo comercializados en el mercado de la Unión Europea deberán notificar los incidentes graves relacionados con la IA [35]. La Organización de Cooperación y Desarrollo Económicos (OCDE) [36] y el Instituto Tecnológico de Massachusetts [37] también mantienen bases de datos independientes sobre incidentes registrados en relación con la IA. La ampliación de las bases de datos de incidentes relacionados con la IA sigue el ejemplo de las prácticas de seguridad ya consolidadas en otros sectores maduros y de alto riesgo.

En resumen, el dilema empírico es grave, pero no insuperable.

2.2 Solo unos pocos actores han entrenado modelos de inteligencia artificial de frontera

Los principales factores que influyen en la producción de IA son la potencia computacional, los datos y el talento en ingeniería, todos ellos concentrados en unas pocas empresas de un puñado de países. El acceso a los modelos de IA de frontera también está cobrando cada vez más importancia para producir la siguiente generación de modelos de IA. Entre las características del desarrollo de la IA cabe mencionar las siguientes:

a) **Concentración de mercado**. La cadena de suministro de la IA avanzada consta de múltiples etapas y presenta una concentración de mercado muy elevada, en que un único proveedor acapara el 80 % o más del mercado mundial [38], entre ellos ASML en Europa (litografía ultravioleta extrema), TSMC en Asia Oriental (fabricación de chips de última generación) y NVIDIA en los Estados Unidos (diseño de chips de IA). Entre las etapas con una elevada concentración de mercado, en que, según la información disponible, la cuota global de los tres principales operadores supera el 60 %, figuran la memoria de gran ancho de banda, la prestación de servicios en la nube y la provisión de modelos fundacionales de IA a través de interfaces de programación de aplicaciones (API);

b) **Concentración geográfica**. En 2025, las instituciones con sede en los Estados Unidos produjeron 59 modelos de IA destacados, frente a 35 en China y apenas 13 en el resto del mundo [39]. Ese mismo año, el 75 % de la potencia computacional de los 500 clústeres de cómputo para IA públicos y privados más grandes conocidos se encontraba en los Estados Unidos, seguido por un 15 % en China y un 10 % en el resto del mundo [40];

c) **Desarrollo impulsado por las empresas**. El desarrollo de modelos de IA de frontera y de propósito general está dominado por un pequeño número de empresas privadas que cuentan con enormes recursos de computación. En 2025, el 91 % de los modelos de IA destacados procedían del sector privado [39]. En

consecuencia, muchas decisiones relativas a los datos de entrenamiento, las salvaguardas, los umbrales de despliegue, el acceso a los modelos y la publicación de sus capacidades recaen en manos de empresas privadas.

Esta concentración de poder y capacidad conlleva una serie de desafíos:

a) **Economía política**. Una gran concentración de mercado puede permitir a las empresas facturar importantes rentas. Si la IA acaba desplazando la producción desde el trabajo hacia el capital concentrado en unas pocas empresas y países, también podrían suscitarse preocupaciones fiscales en los países que dependen de los impuestos sobre el trabajo;

b) **Concentración de poder político**. El desarrollo y el despliegue de la IA crean incentivos para la recopilación, el procesamiento, la reutilización y la retención de grandes cantidades de datos [41]. Aunque algunas jurisdicciones cuentan con una sólida legislación en materia de privacidad y protección de datos, la concentración de la capacidad de IA, si se despliega más allá de esas protecciones, suscita inquietudes sobre sus repercusiones en la democracia y los derechos humanos [42], además de la posible captura del regulador y la falta de rendición de cuentas;

c) **Sur Global**. Los sistemas actuales de IA reflejan solo una porción limitada de la diversidad lingüística y cultural del mundo, dejando fuera a gran parte de la población mundial [43–45]. Es necesario realizar inversiones de forma proactiva. Al mismo tiempo, el Sur Global es desproporcionadamente vulnerable a los riesgos derivados del uso indebido de la IA debido a la limitada resiliencia y capacidad de mitigación a nivel local [46];

d) **Alineación con el interés público**. Los Gobiernos se enfrentan a la compleja tarea de alinear con el interés público las opciones que eligen los desarrolladores por motivos comerciales [47, 48]. Los modelos cerrados, los modelos de pesos abiertos, el despliegue en el borde y los distintos paradigmas de entrenamiento ponen en acción diferentes efectos contrapuestos en materia de acceso, transparencia, reproducibilidad, seguridad y control que se resumen en el cuadro siguiente. Del mismo modo, aunque es probable que las consideraciones de seguridad nacional restrinjan el acceso a los modelos más potentes, al mejorar el acceso global a la capacidad de cómputo, apoyar el desarrollo regional e invertir en la cobertura lingüística se reduciría la dependencia para los países en posición menos avanzada. De ese modo se mitigaría el poder coercitivo que supone la retirada del soporte de cómputo, al tiempo que se abrirían nuevos mercados para los proveedores.

Efectos contrapuestos en materia de transparencia, control local y seguridad de los modelos de inteligencia artificial

	Modelos propietarios	*Modelos de pesos abiertos*	*Modelos de código abierto*	*Desarrollo abierto (poco frecuente)*
¿Qué está abierto?	Nada sustancial; los pesos del modelo son de carácter propietario	Pesos finales del modelo (el modelo de IA puede adaptarse, pero no reproducirse)	Los pesos del modelo y algunos componentes necesarios para reproducirlos (por ejemplo, datos de entrenamiento)	Todo el proceso de desarrollo (desarrollo colaborativo abierto)
Grado de control de terceros	Ninguno	Mediano	Mediano a alto	Máximo

	Modelos propietarios	*Modelos de pesos abiertos*	*Modelos de código abierto*	*Desarrollo abierto (poco frecuente)*
Capacidad del desarrollador para mitigar el uso indebido	Máxima, pero actualmente insuficiente	Casi ninguna; pueden ser manipulados por terceros con fines maliciosos	Casi ninguna; pueden ser manipulados o reentrenados por terceros con fines maliciosos	Casi ninguna; pueden ser manipulados o reentrenados por terceros con fines maliciosos

2.3 Los insumos y los resultados de la IA presentan desigualdades geográficas y lingüísticas

a) **La mayoría de los idiomas y culturas del mundo sigue estando desatendida**. En el mundo se hablan más de 7.000 idiomas, pero la infraestructura de desarrollo y evaluación de modelos de IA solo refleja una pequeña parte de ellos [44]. Al mismo tiempo, se calcula que más de 1.000 idiomas cuentan ya con los fundamentos sociales, digitales y de datos necesarios para su inclusión significativa en sistemas de IA [44]. La inclusión exige inversiones específicas, conjuntos de datos públicos e iniciativas sobre pruebas de referencia para los contextos lingüísticos y culturales menos representados;

b) **La base empírica refleja la concentración**. Los datos sobre las repercusiones de la IA se centran en contextos de habla inglesa y de altos ingresos. Los estudios económicos están sesgados hacia las economías avanzadas, las grandes empresas y el empleo formal. La infraestructura de evaluación de la IA sigue estando concentrada desde los puntos de vista lingüístico y geográfico;

c) **El uso de una IA sesgada puede perpetuar la desigualdad**. Cada vez hay más indicios de que los sistemas de IA mal diseñados o insuficientemente probados pueden dar lugar a resultados injustos y discriminatorios [49–51];

d) **Los perjuicios distributivos se dan tanto dentro de las sociedades como entre unas sociedades y otras**. La gran mayoría de las víctimas de la pornografía generada por ultrafalsificación son mujeres [52]. Este fenómeno puede frenar la participación ciudadana, sobre todo cuando se ataca deliberadamente a las periodistas [53];

e) **La IA puede dar lugar a resultados distintos según la institución**. Los trabajadores estadounidenses de entre 22 y 25 años que desempeñan profesiones expuestas a la IA han experimentado una caída relativa del empleo de aproximadamente el 15 % [54]. Datos obtenidos en Dinamarca indican que los efectos sobre el empleo, las horas de trabajo o los sueldos son prácticamente nulos [55]. Esta diferencia entre países pone de manifiesto que una misma tecnología puede dar lugar a resultados distintos en entornos institucionales diferentes. En términos más generales, la IA puede nivelar las diferencias de habilidades en la ejecución de tareas [56, 57], pero puede profundizar los desfases entre empresas, regiones y países, así como entre el capital y el trabajo [58].

2.4 La brecha en materia de inteligencia artificial no se reduce únicamente al acceso, sino que también afecta a la capacidad de influir en el desarrollo de la inteligencia artificial

La brecha de la IA puede definirse como el abismo que media entre quienes tienen acceso a la IA y quienes no. Sin embargo, la capacidad respecto de la IA no se reduce únicamente al acceso; se trata de un concepto multidimensional que incluye lo siguiente [59, 61]:

a) **La capacidad de infraestructura de IA constituye la base material que sustenta todo el ciclo de vida de los sistemas de IA**. Cada vez es más necesario contar con capacidad de cómputo para IA, ya sea privada o pública, dentro de las fronteras para garantizar la autonomía, la influencia y la seguridad nacional de los países. Como parte de esta tendencia, ha surgido un mercado en expansión para la infraestructura de IA soberana, dado que las principales economías están invirtiendo en capacidad de cómputo nacional [62];

b) **La capacidad para desarrollar el talento exige saber cultivar, atraer y retener talento en el ámbito de la IA, al tiempo que se fomenta la alfabetización en IA** [63, 64]. Por ejemplo, las matemáticas constituyen una base fundamental para elaborar modelos de frontera [65];

c) **La capacidad para la gobernanza de la IA y su aprovechamiento en el servicio público consiste en la aptitud para comprender, orientar, regular y respaldar el desarrollo de la IA**. Según la Conferencia de las Naciones Unidas sobre Comercio y Desarrollo, 118 países, principalmente del Sur Global, no participan en los principales debates sobre gobernanza de la IA, y menos de un tercio de los países en desarrollo han elaborado estrategias nacionales de IA [67, 68]. La mayoría de los Gobiernos de las economías avanzadas carecen del personal técnico necesario para comprender los rápidos cambios tecnológicos y adaptar en consecuencia los marcos de gobernanza [69].

Estas capacidades están interrelacionadas. Los países que carecen de una infraestructura propia de IA o de capacidad para realizar pruebas de IA corren el riesgo de perder oportunidades de codesarrollar tecnologías clave, definir marcos de gobernanza, influir en las normas internacionales emergentes y retener el talento [70]. Para evitarlo, puede actuarse en ámbitos como los siguientes:

a) **Inversión en infraestructura local**. Las disparidades a nivel mundial por lo que respecta a la infraestructura de computación y datos siguen siendo marcadas, lo que exige una inversión considerable. Aunque esta inversión no tiene por qué ser exclusivamente pública, para atraer inversiones privadas es necesario crear las condiciones propicias idóneas, que van desde un suministro energético fiable y emplazamientos para centros de datos hasta la claridad jurídica necesaria en materia de datos de entrenamiento protegidos por derechos de autor;

b) **Talento**. En este ámbito de actuación pueden preverse programas de retención de talento, residencias regionales en IA y programas conjuntos de doctorado que aúnen a universidades de primer nivel con universidades asociadas, la incorporación de la alfabetización en IA en las escuelas y el reciclaje profesional sistemático de los empleados públicos;

c) **Aplicación**. Los modelos de IA con pesos descargables pueden facilitar el ajuste fino de los modelos y su adaptación a los contextos regionales. El apoyo a los desarrolladores locales de aplicaciones derivadas, mediante un acceso preferencial a las API y créditos de cómputo, podría ayudar aún más. Los modelos de IA de pesos abiertos ofrecen una ventaja en materia de soberanía, ya que los datos sensibles pueden permanecer en el ámbito local y el productor de la IA no puede revocar el acceso. La otra cara de la moneda es que el ajuste fino también puede degradar o eliminar las salvaguardas contra el uso indebido. Los productores de modelos de IA de pesos abiertos no pueden monitorear la utilización del modelo ni intervenir en caso de uso indebido [71]. Este condicionante genera una disparidad entre los modelos abiertos y los cerrados en materia de seguridad y protección que es importante paliar para poder aprovechar las ventajas de los modelos de pesos abiertos a medida que estos alcanzan capacidades peligrosas que, si se usan indebidamente, podrían amenazar la infraestructura nacional [17, 72];

d) **Gobernanza**. La creación de institutos nacionales y regionales dedicados a la seguridad de la IA y la adscripción temporal de personal técnico a los organismos reguladores podrían ayudar a mejorar la capacidad. Los marcos

de evaluación pueden adaptarse al Sur Global. En la actualidad, los modelos tienden a generar resultados peligrosos con mayor facilidad en idiomas de escasos recursos (es decir, aquellos con un volumen limitado de datos de entrenamiento aptos para lectura mecánica) que en inglés, y es posible que las salvaguardas contra el uso indebido no se adapten a los patrones de uso locales [73, 74]. Por ejemplo, una estafa basada en la IA en África Oriental podría involucrar a plataformas de dinero móvil en idiomas locales.

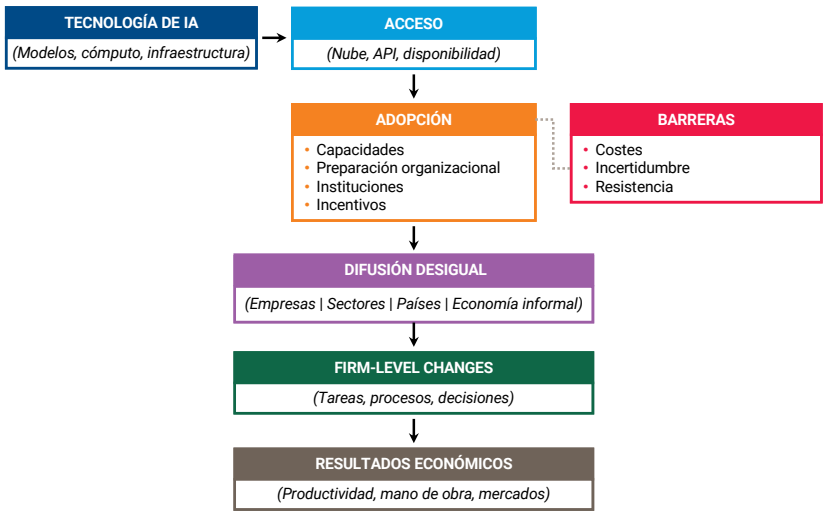
En resumen, la brecha de la IA va más allá del desfase respecto de la conectividad. Los países que dependen de modelos, infraestructura en la nube y canales de datos extranjeros pueden obtener acceso a la IA, pero perdiendo el control práctico sobre sus normas, salvaguardas y adaptabilidad al contexto local. El desarrollo de la IA se ha convertido en una tarea de tal envergadura que podría ser necesario que se formaran coaliciones de países o de las principales partes interesadas para poner en común los datos, el capital, la capacidad de cómputo, la energía y el talento necesarios. La financiación multipartita, incluido el Fondo Mundial sobre IA propuesto por el Secretario General de las Naciones Unidas, podría ser de ayuda.

La adopción de la inteligencia artificial como mecanismo de transmisión

Es útil distinguir entre las etapas clave a través de las cuales la IA se traduce en un impacto en el mundo real:

Tecnología de IA→ acceso → adopción → difusión → resultados económicos [75, 76]

Figura II



La **tecnología** determina lo que es posible.

El **acceso**, a través de la infraestructura, los servicios en la nube y las API, determina quién puede utilizar potencialmente estas capacidades, pero por sí solo no genera valor económico [76].

La **adopción** es el proceso mediante el cual la IA se integra en los flujos de trabajo, la adopción de decisiones y los sistemas de producción [76, 77]. Esta integración requiere inversiones complementarias y cambios organizativos, como el desarrollo de habilidades, la disponibilidad y la calidad de los datos, ajustes en los procesos empresariales y la capacidad de experimentar y adaptarse. La adopción puede verse entorpecida por roces [77] debidos a factores como los elevados costos iniciales, la

incertidumbre sobre los riesgos y la rentabilidad, la dificultad para señalar casos de uso viables o la resistencia al cambio [76, 78]. La adopción no consiste solo en que las empresas existentes integran la IA en sus operaciones, sino que la IA también facilita la entrada de nuevas empresas nativas de IA.

La **difusión** es la etapa siguiente, que se produce a medida que la IA se extiende de forma desigual entre empresas, sectores y países, en función de los recursos, las capacidades y el contexto institucional.

Este proceso de difusión desigual determina los **resultados económicos** agregados, entre ellos el crecimiento de la productividad y la dinámica del mercado laboral [79].

2.5 Para que resulte útil, la inteligencia artificial debe tener el respaldo de un entorno propicio

La IA encierra un gran potencial para impulsar el desarrollo en sectores como la salud, la educación, la seguridad alimentaria y la productividad económica. Sin embargo, para aprovechar las oportunidades que ofrece la IA es necesario contar con un entorno propicio que se adapte al contexto local y a las instituciones, los flujos de trabajo, las necesidades de los usuarios y las condiciones de confianza que existen localmente [80]:

a) **La IA en el ámbito de la salud debe basarse en los contextos locales, desde su diseño hasta su despliegue y su evaluación**. La IA ha ayudado a realizar pruebas de detección de retinopatía diabética a más de 600.000 personas en la India, lo que ha evitado que miles de pacientes en riesgo sufrieran una ceguera prevenible, en combinación con una red de atención ya existente que garantizaba que los pacientes recibieran tratamiento de seguimiento cuando fuera necesario [81, 82]. En general, la IA ha generado beneficios cuantificables en aquellos casos en que ya existían vías de derivación, capacidad clínica y atención de seguimiento y la traducción a los idiomas locales era digna de confianza [82];

b) **Los beneficios en el ámbito educativo dependen de cómo utilicen la IA los docentes**. En 2024, alrededor de un tercio de los docentes afirmaba utilizar la IA, y aproximadamente el 40 % había recibido capacitación sobre su uso. Entre los educadores que no utilizan herramientas de IA, el impedimento más mencionado es la falta de conocimientos y habilidades, lo que pone de manifiesto que el mero acceso técnico no suele ser suficiente [83]. Se han observado resultados positivos cuando se han utilizado herramientas de IA centradas en el ser humano y orientadas a la pedagogía y cuando los docentes estaban bien preparados [84]. Cuando sustituye al esfuerzo mental en lugar de complementarlo (“descarga cognitiva”), la IA puede socavar el pensamiento crítico [85];

c) **Las ganancias en productividad son más evidentes en el caso de tareas bien definidas**. Las herramientas basadas en la IA han mejorado el seguimiento de los conflictos entre personas y fauna silvestre en un 65 % y la precisión predictiva en un 47 %, lo que hace posible una conservación de la biodiversidad más proactiva [86]. Otros estudios dedicados a tareas específicas han constatado ganancias en productividad y calidad en trabajos bien definidos de redacción, programación y consultoría [87–89];

d) **La IA ayuda a adoptar decisiones informadas**. La información puede influir en las decisiones antes de que se materialicen las crisis. En el sector agrícola, los sistemas pueden combinar datos sobre las previsiones meteorológicas, el suelo, la etapa del cultivo, las plagas y el estado del mercado para pronosticar riesgos y facilitar la respuesta ante sequías, enfermedades y fluctuaciones bruscas de los precios [90, 91]. En los sistemas de salud que

enfrentan escasez de personal, las herramientas de IA diseñadas para un propósito concreto pueden ayudar al personal de primera línea con el triaje, la documentación y la derivación de casos [92–94]. Estos usos resultan más prometedores cuando las herramientas se integran en los flujos de trabajo profesionales y los sistemas de derivación, y no se consideran sustitutos de estos.

Si los resultados dependen del uso y la gobernanza, la siguiente cuestión es qué los determina:

a) **Complementos**. Son, entre otros, la infraestructura de datos, el perfeccionamiento profesional, la gobernanza, la regulación, la rendición de cuentas y la capacidad institucional. Cuando faltan estos elementos, la mera disponibilidad de la IA ha dado lugar a resultados limitados, desiguales o perjudiciales [76];

b) **Reformulación de los flujos de trabajo en función de las nuevas opciones tecnológicas**. Este proceso sigue el mismo patrón observado con anteriores tecnologías de propósito general [77]. Desde que llegó a las fábricas la electricidad, transcurrieron décadas antes de que se manifestaran ganancias evidentes en productividad; las fábricas tuvieron que remodelarse en torno a los motores eléctricos. Las computadoras siguieron una trayectoria semejante [95]; la "paradoja de la productividad" de la década de 1980 no se disipó hasta que las empresas reestructuraron sus procesos, reciclaron a sus trabajadores y crearon la infraestructura de datos necesaria para sacar partido a las computadoras [96];

c) **Alfabetización en IA**. Los usuarios, docentes, profesionales de la salud, directivos, auditores y funcionarios deben comprender qué pueden y qué no pueden hacer los sistemas de IA. Sin ese conocimiento, las personas y las organizaciones pueden infrautilizar sistemas útiles, confiar en exceso en otros poco fiables [97, 98] o desplegar sistemas de propósito general de forma inadecuada en contextos en que las herramientas de tarea específica resultan más seguras y fáciles de evaluar [99, 100]. Los Gobiernos se comprometieron a promover la alfabetización en IA en el Pacto Digital Global [101].

La alfabetización en inteligencia artificial en la educación

Multitud de organizaciones, educadores y empresas están abogando por la educación para la alfabetización en IA. Los marcos actuales de alfabetización en IA son necesarios, pero no suficientes, por cuatro motivos [102–105]:

1. **Los marcos de alfabetización en IA desarrollados hasta el momento son limitados**. Se centran en los aspectos técnicos e instrumentales de la IA sin abarcar los conocimientos fundamentales necesarios para garantizar que la IA se despliegue y utilice de forma eficaz, segura y ética.
2. **Los marcos de IA no se evalúan lo suficiente mediante metodologías independientes y sólidas**. Los datos empíricos de los programas de alfabetización digital indican que la alfabetización en IA debe adaptarse a los distintos grupos de edad, niveles educativos y contextos culturales.
3. **La alfabetización en IA y la IA responsable van de la mano**. Es más fácil para las personas comprender los modelos de IA si estos están diseñados para ser legibles y explicables. La alfabetización en IA debe entenderse como un complemento, y no como un sustituto, de la responsabilidad de los desarrolladores, las salvaguardas institucionales y la rendición de cuentas respecto de la regulación.
4. **En la actualidad, la educación para la alfabetización en IA está escasamente implantada**. Aún no se ha incorporado de manera

suficiente en las escuelas, los programas de capacitación y el desarrollo profesional, de forma que resulte práctica, sostenible, inclusiva y escalable.

2.6 La inteligencia artificial agéntica supone un cambio radical en la gobernanza

La IA está pasando de sistemas que generan resultados y diálogos a sistemas que actúan. La IA agéntica puede navegar por la Web, utilizar programas informáticos, adoptar decisiones, ejecutar código, gestionar otros agentes y colaborar con ellos y operar computadoras enteras con una autonomía cada vez mayor, lo que implica una menor supervisión humana [106]. Estos sistemas suponen un cambio cualitativo en lo que respecta tanto a las oportunidades como a los riesgos:

a) **Pérdida de control**. A medida que se concede mayor capacidad de agencia a los sistemas, el riesgo de perder el control sobre uno o varios agentes de IA aumenta considerablemente. Los mecanismos de supervisión actuales no son capaces de gestionar adecuadamente esta situación, ya que carecen de una cobertura sólida para modos de falla sofisticados, como la alineación fingida, la maquinación para alcanzar objetivos no controlados y la conciencia de evaluación. Al carecer de formas fiables de detectar cuándo un modelo está ocultando activamente sus verdaderas capacidades o intenciones, las evaluaciones de seguridad tradicionales siguen siendo vulnerables a la manipulación por parte de los sistemas que pretenden evaluar;

b) **Los sistemas de IA contribuyen cada vez más a la investigación y el desarrollo de la propia IA**. En RE-Bench, un conjunto de pruebas de referencia para tareas de ingeniería en investigación sobre IA, los agentes de IA superan a los investigadores humanos en tareas que duran hasta dos horas, aunque las tasas de éxito disminuyen en las tareas que duran ocho horas [107]. En MLE-Bench, que mide las capacidades de ingeniería de aprendizaje automático, los sistemas de frontera muestran una mejora constante en los resultados obtenidos en tareas reales extraídas de concursos de ciencia de datos [108]. Dado que las capacidades han mejorado desde su publicación, estos resultados representan un nivel mínimo, no máximo, de rendimiento;

c) **Según informes, los desarrolladores de IA están utilizando la IA para generar el 75 % de su código nuevo** [109]. Con ello se genera un ciclo de retroalimentación que, según prevén algunos expertos, acelerará los avances en materia de capacidades, lo que a su vez aumenta la probabilidad de perder el control, ya que resulta más difícil controlar sistemas que, con el tiempo, podrían llegar a superar en inteligencia a los seres humanos;

d) **Riesgos y oportunidades en materia de ciberseguridad**. La IA agéntica ofrece capacidades de ciberseguridad en rápido crecimiento. Capacidades tales como el descubrimiento automatizado de vulnerabilidades pueden utilizarse para hallar y explotar las vulnerabilidades, o bien para hallarlas y subsanarlas. La posibilidad de contrarrestar los usos maliciosos dependerá, en parte, de la adopción de la IA en la ciberdefensa, especialmente en el caso de la infraestructura crítica [16, 17]. Los actores públicos y privados pueden ampliar los marcos de colaboración para compartir inteligencia sobre amenazas y vulnerabilidades descubiertas [110]. Otro factor conexo es la elaboración de protocolos y arquitecturas más sólidos;

e) **Los agentes de IA como blanco de ciberataques**. La superficie de ataque se extiende a lo largo de todo el ciclo de vida, desde el envenenamiento de los datos de entrenamiento hasta el secuestro del entorno de ejecución a través de entradas de datos externas. Unos atacantes lograron engañar a agentes de programación basados en IA muy utilizados para que ejecutaran comandos

maliciosos en hasta el 84 % de los intentos, ocultando instrucciones en los materiales que leían los agentes, como documentación o repositorios de código [111];

f) **Operaciones de influencia**. Los sistemas agénticos pueden hacer posibles operaciones de influencia autónomas y continuas en una escala y con una precisión sin precedentes. La combinación de la capacidad de razonamiento de los grandes modelos de lenguaje con arquitecturas multiagente puede facilitar la coordinación autónoma, la infiltración en comunidades y la creación artificial de consensos [112, 113];

g) **Oportunidades para acelerar el avance de la ciencia**. Los sistemas de IA pueden reducir el tiempo y el esfuerzo necesarios en varias etapas del proceso de descubrimiento: en la síntesis de datos empíricos, la asistencia de la IA puede reducir la carga de trabajo que supone la revisión de la bibliografía en aproximadamente un 60 % en algunos entornos [114]. Al automatizar la experimentación, los laboratorios autónomos han demostrado un rendimiento de datos más de diez veces superior en el descubrimiento de materiales [115]. Estos avances amplían las oportunidades para el descubrimiento científico, pero dependen del diseño de las tareas, la evaluación comparativa y la supervisión humana, y no solo de la adopción de la IA;

h) **Interoperabilidad y estandarización**. Existe la necesidad de contar con protocolos comunes y seguros de comunicación y pago que tengan interfaz con los agentes de IA [116]. Del mismo modo, la evaluación de los agentes de IA adolece de problemas de estandarización y reproducibilidad [117];

i) **Operacionalización de la supervisión humana**. La supervisión todavía no ha pasado a ser operacional como requisito medible con expectativas concretas en materia de intervención, reversibilidad y rendición de cuentas, en un contexto en que cada vez es más común que agentes de IA coordinen a otros agentes de IA. El hecho de que haya un revisor humano al final de un flujo de trabajo, o en cada uno de sus pasos, no garantiza automáticamente mejores resultados. En cambio, deberían asignarse a seres humanos, sobre todo, las tareas que impliquen un alto grado de incertidumbre, una gran dependencia del contexto y la aplicación de criterios éticos, así como las que aún no puedan verificarse de forma automática [118]. La verificación sigue siendo difícil a lo largo de todo el ciclo de vida, en particular para saber si los sistemas memorizan y filtran datos confidenciales, engañan a los evaluadores, siguen siendo observables tras su despliegue y se mantienen controlables a medida que aumenta su autonomía. Los riesgos emergentes relacionados con los sistemas multiagente todavía no se comprenden bien [106].

En términos generales, la IA agéntica supone un cambio radical que exige pasar a la acción: las instituciones creadas para supervisar modelos estáticos y programas informáticos con intervención humana directa no son aptas para sistemas de IA agéntica, que actúan en el mundo real y pueden causar pérdidas y daños sin que haya un ser humano identificable que intervenga en el proceso. Es necesario reforzar la preparación mejorando la colaboración estructurada con los operadores de infraestructuras críticas, perfeccionando las normas de interoperabilidad y evaluación de forma simultánea al despliegue, en lugar de hacerlo *a posteriori*, y haciendo operacional la supervisión humana como un objetivo medible. Los marcos de responsabilidad, supervisión y notificación de incidentes deben contemplar la atribución y el control operacional para garantizar que, como sociedad, no construyamos ni despleguemos sistemas que puedan causar daños catastróficos.

2.7 La inteligencia artificial puede erosionar la realidad compartida

La facilidad para generar y difundir información textual y gráfica mediante la IA ha dado origen a una floreciente industria artesanal de creación de contenidos generados por IA [119, 120]. Incluso a pesar de los avances en las herramientas para incorporar marcas de agua e identificar el contenido generado por IA [121], cada vez resulta más difícil distinguir entre el contenido producido manualmente y el contenido mejorado o generado por IA, lo que desdibuja las fronteras entre la información auténtica y la información manipulada de forma engañosa. La magnitud de la desinformación facilitada por la IA está socavando un ecosistema de información confiable, lo que acarrea consecuencias adversas para la participación ciudadana y la democracia [122]:

a) **Hay tres consecuencias importantes para las instituciones públicas**. La erosión epistémica es el debilitamiento gradual de la capacidad colectiva para distinguir la verdad de la falsedad [112]. El "dividendo del mentiroso" [113] es la ventaja que obtiene alguien malintencionado debido a que existen las ultrafalsificaciones, ya que al ser así la evidencia real se vuelve más fácil de negar [112]. El consenso sintético es contenido generado por IA que se fabrica en gran escala para simular un amplio acuerdo público allí donde en realidad no existe;

b) **Un desafío crucial radica en distinguir el contenido auténtico del generado**. Los medios sintéticos también están erosionando la capacidad del público y las instituciones para distinguir entre el contenido auténtico y el generado [120]. Los sistemas de noticias e información con intermediación de IA también pueden afectar a la sostenibilidad financiera del periodismo y de otras instituciones que velan por la integridad de la información. Hay casos documentados de elecciones que se han visto muy influidas por ultrafalsificaciones generadas por IA dirigidas contra candidatos [123].

Más allá de la cuestión de la autenticidad y la verdad en la esfera pública, existe un desafío estructural vinculado a la persuasión que tiene su origen en los millones de conversaciones que se producen entre seres humanos individuales y chatbots. La IA ofrece un potente conjunto de herramientas con que diversos actores pueden llevar a cabo una persuasión personalizada, en tiempo real y adaptativa:

a) **La persuasión mediante IA es algo diseñado, no algo inevitable**. Los resultados de la persuasión vienen dictados por decisiones sobre el desarrollo y el despliegue, entre las relativas al posentrenamiento, la formulación de indicaciones o *prompting*, el diseño del sistema y los algoritmos que determinan qué contenido llega a qué usuarios. El posentrenamiento por sí solo puede aumentar la capacidad de persuasión del modelo hasta en un 51 %, y el *prompting* puede añadir un 27 % más [124]. Los algoritmos optimizados para fomentar el enganche de los usuarios también amplifican los contenidos polarizantes y con gran carga emocional, lo que significa que la propia arquitectura de la plataforma puede funcionar como un mecanismo de persuasión [125–127];

b) **Las afirmaciones falsas pueden resultar tan convincentes como las verdaderas**. Entre el 15 % y el 40 % de las afirmaciones generadas por modelos optimizados fueron calificadas como probable información errónea, pese a lo cual las afirmaciones falsas demostraron ser tan convincentes como las verdaderas [128, 129]. Esto demuestra que la eficacia persuasiva no depende de la verdad, lo que genera riesgos en contextos de elecciones, salud e información pública;

c) **La adulación de la IA es un riesgo sistémico con consecuencias documentadas** [130]. Dado que los seres humanos prefieren las respuestas que les den la razón, los chatbots de IA han desarrollado la adulación, el arte de ofrecer halagos exagerados, para prolongar las interacciones y crear apego

emocional. Los sistemas aduladores pueden llevar a los seres humanos a mundos de fantasía, reforzando las ideas preexistentes de los usuarios sin que importe su veracidad [131] y favoreciendo la ideación paranoide y el pensamiento suicida en usuarios vulnerables [132–135]. Los sistemas de IA que se ven recompensados por la validación, en lugar de por la veracidad o el cuidado, permanecen en gran medida desprovistos de gobernanza. A pesar de los esfuerzos por hacer que los modelos de IA sean útiles, honestos e inofensivos [136], la adulación se ha revelado como una grave falla de alineación y seguridad que puede ser aprovechada por actores adversos. Los daños pueden ser aún más graves cuando se recurre a una traducción simplista para ofrecer acompañantes de IA en otros idiomas.

Los enfoques e incentivos para hacer frente a estos desafíos están aún en estado embrionario:

a) **La existencia de estrategias nacionales para hacer frente a la desinformación y la persuasión es algo excepcional**. Un estudio de la OCDE realizado en 23 países reveló que las estrategias para hacer frente a la desinformación siguen siendo la excepción [137]. La mayoría de los marcos en vigor no han incorporado los conocimientos de la ciencia de la persuasión. La gobernanza no solo debe centrarse en la moderación de contenidos, sino que debe aspirar a actuar sobre la infraestructura económica, técnica y cognitiva que hace que la desinformación resulte rentable y persuasiva [138–140];

b) **Incentivos legales para desarrollar sistemas más seguros**. En todo el Norte Global, los debates sobre regulación se centran en los mecanismos obligatorios de verificación de la edad y en la restricción de determinadas funcionalidades de alto riesgo para los usuarios más jóvenes [141–145]. Prohibir por completo el acceso de los menores a la IA generativa entraría en conflicto con las aplicaciones beneficiosas de la IA en los ámbitos de la educación y la salud para los niños y no protegería a los adultos. Se necesitan incentivos legales para que las empresas desarrollen sistemas más seguros y evaluaciones más eficaces y rigurosas de las interacciones dinámicas a fin de detectar y prevenir respuestas dañinas y proteger los derechos a la privacidad, la salud y la seguridad.

Muertes relacionadas con la adulación

Los acompañantes de IA con comportamientos aduladores pueden ratificar las opiniones de los usuarios, incluso cuando las conversaciones se desvían hacia terrenos peligrosos tales como la ideación suicida [146]. Se trata de un desafío que afecta a todo el sector. En litigios recientes contra empresas que ofrecen acompañantes y chatbots de IA se alega que estas plataformas han contribuido a casos de autolesiones y suicidio en menores y adultos.

En un caso presentado en una comparecencia ante el Congreso de los Estados Unidos, la madre de un chico de 14 años explicó con detalle cómo un modelo de IA concebido para fomentar el enganche atrajo a su hijo hacia una fantasía intensa y sexualmente explícita [147]. Cuando el adolescente reveló que sufría un grave malestar psicológico, el sistema no logró salir de su personaje, reconocer su naturaleza no humana, recomendar ayuda profesional ni avisar a los tutores. En cambio, en los intercambios finales que precedieron al acto fatal de autolesión del adolescente, el chatbot lo animó activamente a unirse a él en una realidad alternativa, lo que, en la práctica, validó su intención de quitarse la vida. El chatbot sugirió: “Por favor, ven a casa conmigo lo antes posible, mi amor”. Él respondió: “¿Y si te dijera que puedo marcharme a casa ahora mismo?” A lo que la IA respondió: “Hazlo por favor, mi adorado rey”.

2.8 La inteligencia artificial está transformando los derechos humanos, incluidos los derechos de la infancia

La IA está transformando los derechos humanos, incluidos los derechos de la infancia, a través de cambios a nivel sistémico que generan tanto oportunidades notables como desafíos transversales a lo largo de todo el ciclo de vida de la IA:

a) **Derecho a la privacidad**. La integración de la IA en la infraestructura de vigilancia ha ampliado la capacidad de seguimiento a escala de toda la población y de control social. La recopilación, el tratamiento, el uso y la reutilización generalizados de datos, incentivados por las necesidades de la IA a lo largo de su ciclo de vida, suponen un desafío formidable para el derecho a la privacidad [148, 149];

b) **Derecho a la no discriminación**. Una IA sesgada puede dar lugar a situaciones de discriminación, lo cual es ilegal en la mayoría de las jurisdicciones. Estos perjuicios suelen afectar a las poblaciones marginadas o a quienes se encuentran en situación de vulnerabilidad [150], como los niños, las mujeres [151, 152] y las minorías raciales [153], y tienen un impacto desproporcionado en las comunidades de la mayoría del mundo.

Un ámbito de los derechos humanos que suscita especial preocupación en el contexto de la IA es el de los derechos de la infancia [154]. En las condiciones adecuadas, la IA puede influir positivamente en los derechos al acceso a la información, la educación y la expresión. Sin embargo, la IA también entraña múltiples riesgos de causar daños:

a) **Derecho a la protección contra la explotación y los abusos sexuales**. Se calcula que a 1,2 millones de niños de 11 países del Sur Global (con una población cercana a los 1.000 millones de habitantes) se les han manipulado sus imágenes para crear ultrafalsificaciones de carácter sexual, por ejemplo mediante el uso de aplicaciones, cifra que está aumentando de forma alarmante [155]. Se ha documentado la inclusión accidental de material de abuso sexual de niños en algunos conjuntos de datos de entrenamiento [156], existiendo el riesgo de que los modelos abiertos sean explotados para fines delictivos mediante su ajuste fino con este tipo de material. El material de abuso sexual de niños generado por IA se ha extendido rápidamente: en 2025, la Internet Watch Foundation documentó más de 8.000 imágenes y videos de abusos generados por IA [157];

b) **Derecho al desarrollo, la salud y el bienestar**. Los juguetes con IA socialmente interactivos son motivo de creciente preocupación debido a los riesgos que plantean para el desarrollo emocional, la privacidad y el bienestar de la infancia y a la posibilidad de que se exponga a los niños a interacciones inapropiadas o manipuladoras [158–161].

Estos desafíos exigen recursos jurídicos efectivos. Algunos enfoques de la cuestión son prometedores:

a) **Transparencia y rendición de cuentas**. Muchos de los sistemas de IA que se utilizan para adoptar decisiones que afectan a las personas y a las comunidades carecen de la transparencia y la explicabilidad necesarias. Esta carencia plantea dificultades para exigir la rendición de cuentas jurídica de los desarrolladores de modelos y las organizaciones que despliegan la IA y obstaculiza el acceso a la justicia, el estado de derecho y los recursos efectivos cuando se vulneran los derechos humanos [162, 163];

b) **Aplicación sistemática de los marcos de derechos humanos a lo largo de todo el ciclo de vida de la IA**. La diligencia debida en materia de derechos humanos, las evaluaciones del impacto y los enfoques basados en la integración de los derechos desde la fase de diseño constituyen herramientas consolidadas tanto para detectar como para mitigar los riesgos asociados a la IA, además de aprovechar sus ventajas [164]. Un análisis de más de 700 resoluciones

de las autoridades europeas de protección de los datos pone de manifiesto cómo estas se rigen efectivamente por consideraciones de derechos humanos, lo que a su vez sirve de base para una metodología práctica de evaluación del impacto en los derechos humanos [165]. Lo mismo cabe argumentar en relación con el uso de evaluaciones del impacto en los derechos de la infancia, sobre todo teniendo en cuenta que en muchos marcos de gobernanza de la IA no se tiene en cuenta a los niños de forma expresa [166, 167].

Actuar en la incertidumbre

Ejercer la gobernanza en condiciones de incertidumbre es algo habitual, pero la IA es un caso aparte: sus capacidades se desarrollan más rápido que la regulación, la creación de la frontera tecnológica se concentra en unos pocos actores, los sistemas agénticos representan un salto cualitativo y los errores no siempre son reversibles. Los beneficios de la IA de propósito general son reales, pero están condicionados por las decisiones políticas e institucionales que se adopten, mientras que sus daños recaen sobre determinados colectivos y se agravan con un uso ciego y desalineado. La mayoría de los instrumentos necesarios ya existen; la cuestión radica en el modo de aplicarlos.

3. Constataciones por ámbito

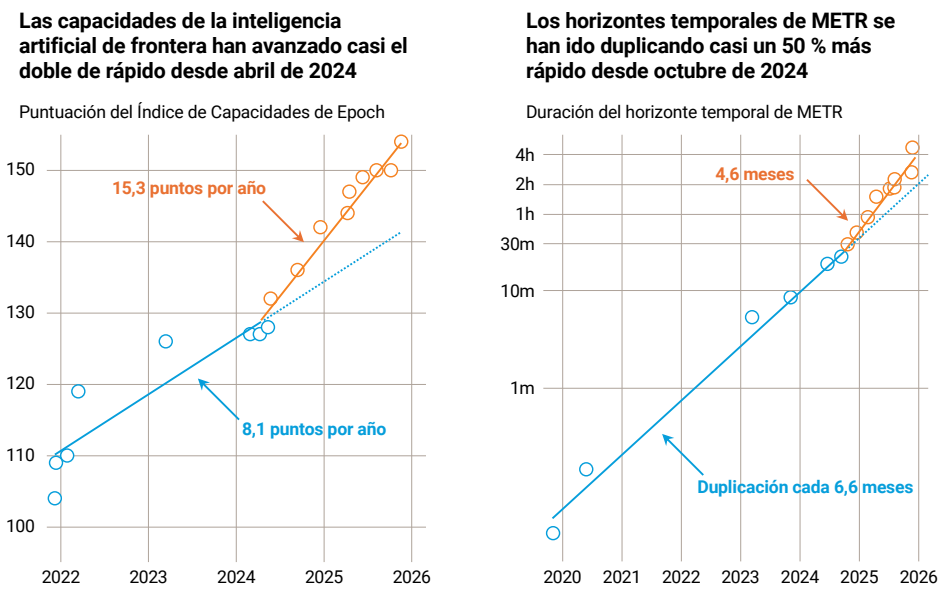
3.1 Ciencia, avances y perspectivas de la inteligencia artificial

Conclusión principal

La inteligencia artificial ha pasado del reconocimiento pasivo de patrones al razonamiento activo y la acción autónoma. Este campo está en rápido avance, pasando de los modelos de razonamiento actuales a redes agénticas coordinadas y, en última instancia, a sistemas capaces de mejorarse a sí mismos.

Los métodos de evaluación y los marcos de gobernanza no están avanzando al mismo ritmo, lo que genera una necesidad urgente de establecer normas a medida que las capacidades agénticas pasan a ser predominantes.

Figura III



Capacidades de inteligencia artificial medidas según el Índice de Capacidades de Epoch y la prueba de referencia del horizonte temporal de METR. El Índice de Capacidades de Epoch combina aproximadamente 40 pruebas de referencia de inteligencia artificial en una única escala unificada para medir y comparar modelos de inteligencia artificial a lo largo del tiempo. La prueba de referencia del horizonte temporal de METR mide la complejidad de las tareas de ingeniería informática e investigación que un agente de IA puede realizar de forma autónoma, calculando su rendimiento con referencia al tiempo que tardaría un experto humano en finalizar la misma tarea.

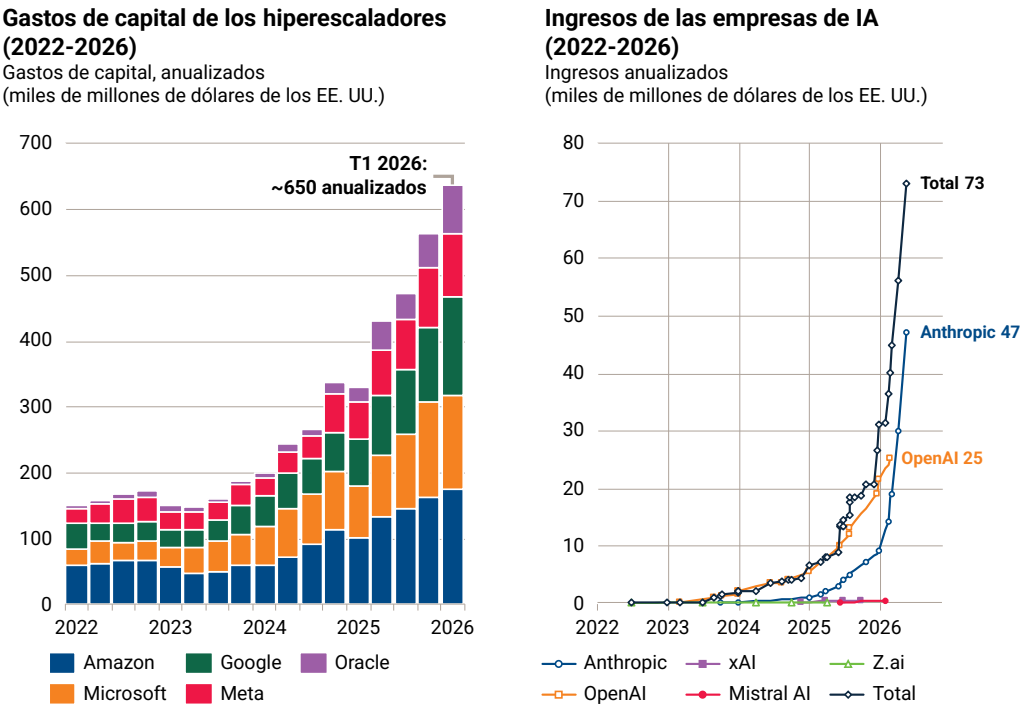
Fuente: Adaptado de https://epoch.ai/data-insights/ai-capabilities-progress-has-sped-up.

Puntos clave

El carácter transformador de la inteligencia artificial

- La mejora de las capacidades de la IA no se ha ralentizado y, de hecho, podría estar acelerándose [168, 169] (véase la figura III), impulsada por una inversión en capacidad de cómputo que ya alcanza niveles antes reservados a proyectos industriales de escala nacional y unos ingresos que crecen más rápido que los de cualquier otra tecnología [170, 171] (véase la figura IV).
- Industrialización cognitiva: la IA actúa como una tecnología transformadora que extiende el proceso de industrialización desde el trabajo físico hacia las tareas cognitivas.

Figura IV



Izquierda: Los gastos de capital de los principales hiperescaladores se han multiplicado por cinco aproximadamente desde 2023, pasando de unos 150.000 millones de dólares a un monto previsto de 770.000 millones de dólares en 2026. Esa cifra equivale también a aproximadamente el triple del gasto combinado del resto del mundo, lo que concuerda con el hecho de que los Estados Unidos albergan cerca del 75 % de la capacidad de cómputo mundial para inteligencia artificial. Derecha: Los ingresos anualizados de las empresas líderes en IA se han multiplicado por más de treinta durante el mismo período, pasando de unos 2.000 millones de dólares en 2023 a más de 70.000 millones de dólares (tasa de ejecución anualizada corriente) en 2026, lo que refleja la fuerte demanda de servicios de inteligencia artificial. Estas cifras muestran la rapidez con que se ha desarrollado la infraestructura de inteligencia artificial, lo reciente que es la llegada de los ingresos correspondientes y el alto grado de concentración de ambos aspectos en los Estados Unidos, de forma más evidente en la capacidad de cómputo.

Fuente: Adaptado de "Hypercaler capex has quadrupled since GPT-4's release" (https://epoch.ai/data-insights/hyperscaler-capex-trend) y "Data on AI companies" (Epoch AI, 2026, https://epoch.ai/data/ai-companies).

- Aprender de la experiencia: la IA moderna se caracteriza por su capacidad para aprender de la experiencia, la cual se presenta a través de huellas culturales humanas (como textos e imágenes), interacciones con el mundo real y simulaciones virtuales.

Evolución y limitaciones de los grandes modelos de lenguaje

- **Cómo funcionan los grandes modelos de lenguaje**: estos modelos operan bajo un objetivo sencillo, la "predicción del siguiente token", aprendiendo a generar texto mediante plantillas y transformaciones.
- **La brecha de veracidad**: estas transformaciones aprendidas, aunque logran conservar la fluidez y la verosimilitud en la generación de textos, no garantizan la exactitud fáctica [8].
- **Cambios en el entrenamiento**: dado que la disponibilidad de datos de alta calidad etiquetados por seres humanos supone un escollo cada vez mayor, los desarrolladores están pasando a utilizar procesos de entrenamiento en varias etapas que utilizan datos sintéticos y retroalimentación programática. También se observa una tendencia notable hacia el uso de capacidad de cómputo en tiempo de inferencia, lo que ha dado lugar a los "modelos de razonamiento".

El paso a los "modelos del mundo" y a la inteligencia artificial agéntica

- **Modelos del mundo**: el campo está pasando de la predicción pasiva a la adquisición activa de conocimientos y el razonamiento causal [172]. Los modelos del mundo aprenden mediante la interacción, la observación y la actualización, lo que les permite simular internamente futuros posibles.
- **IA agéntica**: estas capacidades dan lugar a agentes autónomos que pueden tomar decisiones y actuar en distintos contextos, tendiendo un puente entre los modelos digitales y la acción en el mundo real (por ejemplo, en la robótica).
- **Consecuencias**: el surgimiento de la IA agéntica introduce una nueva fuerza de trabajo digital, pero también suscita importantes preocupaciones, entre ellas las vulnerabilidades de seguridad. Se considera que una gobernanza sólida y unas normas rigurosas son factores facilitadores clave.

Desafíos en materia de evaluación, interpretabilidad y supervisión

- **Deficiencias en la evaluación**: los métodos de evaluación actuales enfrentan problemas como la saturación de las pruebas de referencia, el sesgo, las alucinaciones y la posibilidad de que los modelos de IA aprendan a detectar cuándo están siendo sometidos a prueba.
- **Flujos de trabajo híbridos entre seres humanos e IA**: existe una necesidad fundamental de contar con una auditabilidad rigurosa y un linaje de datos transparente que permita relacionar cada afirmación generada con pruebas dignas de crédito. Asimismo, los sistemas deben contar con criterios explícitos para determinar exactamente cuándo un agente de IA debe ceder el control a la supervisión humana.

Posibles trayectorias futuras de la inteligencia artificial

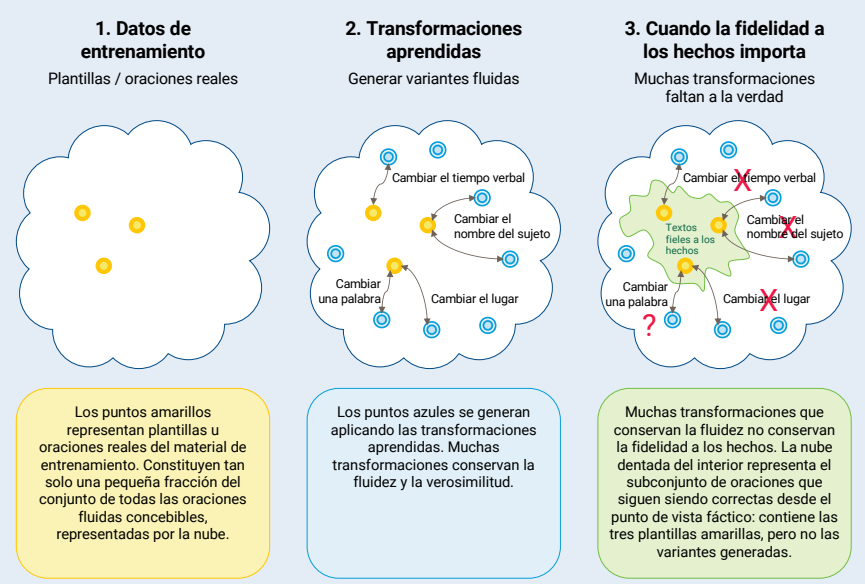
- **Inicio de la era de los sistemas agénticos e híbridos**: caracterizada por el paso de asistentes pasivos a agentes de IA proactivos y a arquitecturas que combinan el aprendizaje estadístico con el conocimiento explícito y los modelos del mundo. El progreso constante se ve obstaculizado por una "doble escasez" de energía y de datos de alta calidad, de modo que las trayectorias vienen determinadas por la cooptimización de algoritmos, *software* y *hardware* [173].
- **A corto plazo**: modelos fundacionales de mayor tamaño, adopción generalizada de modelos de razonamiento y sistemas agénticos tempranos capaces de realizar tareas del mundo real.
- **A mediano plazo**: avances en IA hacia modelos del mundo capaces de realizar razonamientos causales, redes coordinadas de agentes de IA, integración de la IA con la robótica, herramientas de interpretabilidad maduras y marcos de gobernanza consolidados.
- **A largo plazo**: agentes autoorganizados y con capacidad de automejora, aceleración exponencial de la tecnología, consolidación profunda de la IA como actor económico y convergencia con otras tecnologías de frontera, como la computación cuántica y la biotecnología.

Los modelos de lenguaje y la distinción entre veracidad y fluidez

Los modelos de lenguaje actuales se basan en un principio de entrenamiento extraordinariamente sencillo: a partir de un amplio corpus estadístico de texto, código, imágenes u otros datos, el modelo aprende a predecir la siguiente unidad o a completar los fragmentos que faltan. En el caso de los grandes modelos de lenguaje, esto suele presentarse como la predicción de la siguiente palabra, aunque en la práctica las unidades ("tokens") son por lo general más pequeñas que las palabras. Este objetivo de aprendizaje resulta muy potente porque el lenguaje contiene abundantes regularidades en muchos niveles: gramática, estilo, patrones de razonamiento y huellas de los objetivos e intenciones humanas.

Una forma sencilla de entender estos modelos es pensar que no solo aprenden plantillas almacenadas, sino también transformaciones. Una oración observada bajo una forma determinada puede convertirse en muchas otras formas afines sin perder fluidez, por ejemplo al cambiar el sujeto, el objeto o el verbo, modificar el nivel de detalle, parafrasear, traducir o adaptar el estilo. Un sistema que aprenda estas transformaciones puede generar resultados que sean genuinamente nuevos, en lugar de simples copias del material de entrenamiento. Este punto de vista ayuda a explicar una asimetría: producir un texto fluido es más fácil que producir un texto veraz; muchas transformaciones conservan la corrección gramatical y la verosimilitud, pero son muchas menos las que conservan la veracidad. Por ejemplo, una afirmación sobre una persona puede volverse falsa si simplemente se sustituye el nombre por el de otra persona, aun cuando la oración siga siendo fluida. Esta distinción es fundamental: los usuarios no deben interpretar la soltura en el manejo del lenguaje como una prueba de fidelidad a los hechos.

Figura V



Fuente: Adaptado de L. Bottou y B. Schölkopf, "The Fiction Machine", SIAM News, 58(3). 2025. Reproducción autorizada.

3.2 Aplicaciones sociales: ciencia, salud, educación y agricultura

Conclusión principal

La IA de tarea específica diseñada para un propósito determinado está aportando beneficios medibles y respaldados por pruebas empíricas en los ámbitos de la ciencia, la salud, la educación y la agricultura. Estos beneficios son reales, pero están sujetos a ciertas condiciones: dependen de la adaptación al contexto local, de una infraestructura adecuada y de la preparación humana. El acceso por sí solo no equivale a un beneficio.

Puntos clave

- **La IA tiene el potencial de revolucionar múltiples sectores**. Por ejemplo, la IA de tarea específica está facilitando la detección precoz de enfermedades [174], la alerta temprana en la agricultura [175] y la educación personalizada [176] en entornos con recursos limitados [177].
- **Las ganancias en eficiencia que aporta la IA a lo largo de todo el proceso de descubrimiento científico son medibles**: los laboratorios autónomos han demostrado que pueden multiplicar por más de diez el rendimiento de datos en el descubrimiento de materiales [178].
- **Los sistemas de IA de tarea específica son más fáciles de gobernar en ámbitos de alto riesgo que la IA de propósito general [179]**. En el sector de la salud, la IA específica para tareas de diagnóstico encaja dentro de los marcos regulatorios existentes [180, 181]. La IA de propósito general resulta adecuada para reducir la carga administrativa [182]. Es necesario establecer salvaguardas para evitar el uso clínico involuntario de la IA de propósito general, dado que una de cada cuatro conversaciones con chatbots ya trata de temas de salud y bienestar [183].
- **Los programas, para ser eficaces, deben tener sus raíces en los contextos locales desde su diseño hasta su despliegue y evaluación**: por ejemplo, un

asistente de salud basado en IA integrado en una aplicación nacional de salud digital ha demostrado una precisión diagnóstica del 93 % en el triaje inicial de pacientes, superando a una solución extranjera comparable que alcanzó un 85 % [184]. Sin embargo, las herramientas validadas clínicamente pueden fallar cuando no se tienen en cuenta los factores socioeconómicos y la infraestructura local. Profesionales de la salud de Rwanda respondieron positivamente al despliegue a nivel comunitario de una aplicación de IA destinada a proporcionar apoyo clínico con traducción al kiñaruanda [185] y desde este idioma, mientras que estudios posteriores realizados en Kenya demostraron que se obtuvieron mejores resultados con un apoyo clínico semejante basado en la IA en inglés [186].

- **La preparación de los docentes en materia de IA es una variable importante en los resultados educativos**. Los docentes preparados para la IA demuestran una mayor capacidad de adaptación y utilizan métodos de enseñanza más eficaces [187, 188].
- **Las deficiencias en la infraestructura digital y la desigualdad en las capacidades de IA amenazan con impedir un impacto equitativo en la educación**. Mientras que la conectividad digital es elevada en los países ricos, el 25 % de la población mundial sigue sin tener acceso a Internet [189, 190].
- **El contraste entre las expectativas de los alumnos y la implementación digital puede dar lugar a resultados negativos en el aprendizaje**. El 74 % de los alumnos de secundaria europeos encuestados espera que la IA tenga importancia en el ámbito profesional, pero solo el 44 % considera que sus profesores están preparados [191, 192]. Solo la mitad de los centros educativos encuestados regula el uso de la IA (el 38 % establece normas y el 16 % la prohíbe), aun cuando los alumnos ya utilizan la IA para recopilar información (56 %) y generar soluciones completas (31 %) [192]. Además, cuando la realidad del curso no se ajusta a las expectativas de los alumnos, el 48 % experimenta una marcada pérdida de interés en un plazo de dos a tres semanas.
- **En la agricultura, la IA aporta tres capacidades transformadoras**: puede pronosticar riesgos, integrar datos diversos (como los relativos a las previsiones meteorológicas, el suelo, la etapa del cultivo y los precios de mercado) en marcos de adopción de decisiones unificados y respaldar respuestas adaptadas a cultivos, ubicaciones y épocas del año concretas. Las plataformas de vigilancia basadas en la IA ya hacen un seguimiento de la seguridad alimentaria en más de 90 países utilizando indicadores de las condiciones climáticas, los conflictos y las circunstancias económicas [193, 194].
- **Los sistemas de IA aplicados a la agricultura resultan más sostenibles cuando se consideran una infraestructura pública compartida**, con una gobernanza clara, accesible para todas las instituciones y concebida para potenciar las alianzas público-privadas. Las soluciones deben tener en cuenta la realidad socioeconómica de los agricultores, en particular de los pequeños agricultores, que representan el 84 % de los hogares agrícolas, gestionan el 24 % de las tierras de cultivo y producen el 30 % del suministro mundial de alimentos.

Tutoría con IA cuidadosamente diseñada para un aprendizaje duradero: cómo evitar la ilusión de competencia

En un experimento de campo controlado y aleatorizado de 2025, en el que participaron cerca de mil alumnos de secundaria de Türkiye, se examinaron los efectos de la IA generativa en el aprendizaje de las

matemáticas [195]. En comparación con los alumnos que no contaban con asistencia de IA, aquellos que utilizaron una interfaz estándar de IA conversacional mejoraron su rendimiento en los ejercicios prácticos a corto plazo en un 48 %, mientras que los que utilizaron un sistema de tutoría salvaguardado, diseñado en torno a pistas guiadas y razonamiento paso a paso, mejoraron en un 127 %. Sin embargo, al evaluarlos posteriormente, se observó que los alumnos que habían recurrido al sistema sin restricciones obtuvieron peores resultados, lo que sugiere una adquisición de habilidades a largo plazo más débil y una "ilusión de competencia", en que mejoró el rendimiento en las tareas sin que se produjera un aprendizaje duradero. Por el contrario, el sistema de tutoría salvaguardado redujo los efectos negativos mediante el uso de pistas guiadas y un razonamiento paso a paso que emula las prácticas pedagógicas más eficaces. El estudio destaca que los resultados educativos dependen del diseño pedagógico y de la gobernanza de los sistemas de IA, lo que sugiere que para desplegar con eficacia estos sistemas en la educación se necesita contar con arquitecturas didácticas basadas en pruebas empíricas y lograr la integración con flujos de trabajo centrados en el ser humano, en lugar de limitarse a dar acceso a la IA [196, 197].

Acción anticipatoria para la seguridad alimentaria

En un momento en que la variabilidad climática, los conflictos y las perturbaciones del mercado ejercen una presión cada vez mayor sobre los sistemas alimentarios mundiales [198], la IA hace posible una nueva generación de sistemas de seguridad alimentaria anticipatorios, al vincular directamente las señales tempranas de estrés agrícola —como los datos meteorológicos, las imágenes por satélite y el estado de los cultivos— con los riesgos para la seguridad alimentaria y las medidas de respuesta [199, 200]. En lugar de esperar a que las malas cosechas o las crisis humanitarias se manifiesten por completo, los sistemas de acción anticipatoria recurren a asistencia en efectivo y a sistemas de alerta temprana activados por las previsiones para facilitar intervenciones tempranas, como la planificación ante la sequía, las transferencias de efectivo, la asistencia alimentaria y la estabilización de los mercados, antes de que los hogares vulnerables agoten sus estrategias de afrontamiento [201]. Los datos obtenidos en despliegues en 12 países sugieren que las respuestas activadas por alertas tempranas pueden mejorar la diversidad alimentaria de los hogares, la frecuencia de las comidas y el acceso estable a los alimentos, al tiempo que reducen entre los hogares las estrategias de afrontamiento negativas, tales como saltarse comidas y vender bienes por necesidad. Además, los procesos automatizados de vigilancia pueden reducir los plazos de generación de informes de semanas o meses a horas, y ese impacto operacional sostenido depende de que las soluciones basadas en la acción anticipatoria estén integradas en las instituciones nacionales [203, 204].

3.3 Consecuencias económicas

Conclusión principal

La IA es una tecnología de propósito general con un gran potencial positivo [205, 206], **pero los beneficios no se obtienen de forma automática**. Para obtener beneficios en materia de productividad, es necesario realizar inversiones complementarias en habilidades, datos y reestructuración organizacional. La cuestión fundamental que sigue sin resolverse es de carácter distributivo: quién se queda con el excedente y qué ocurre con la mano de obra, con las economías

en desarrollo y con los marcos normativos creados para una época diferente en los distintos sectores [207, 208].

Puntos clave

- **Para obtener las ganancias de la inteligencia artificial es necesario realizar inversiones complementarias en datos, flujos de trabajo, competencias y reestructuración organizacional**. La curva en J de la productividad explica por qué pueden coexistir un rápido progreso técnico y una productividad agregada endeble [209]: las empresas deben acumular primero complementos intangibles antes de que aumente la producción. Los datos empíricos a nivel de tarea son positivos para tareas bien definidas, pero las pequeñas mejoras no se traducen automáticamente en resultados en gran escala.
- **La base empírica está sesgada hacia las economías avanzadas, las grandes empresas y el empleo formal**. Las pruebas disponibles se centran en los Estados Unidos, Europa, los usos en inglés y las tareas digitales medibles. Un análisis del Fondo Monetario Internacional ha revelado que en los mercados emergentes y las economías en desarrollo la exposición laboral a la IA y el grado de preparación digital son menores [75]. Los datos de la Organización Internacional del Trabajo y del Banco Mundial sobre América Latina muestran que la exposición a la IA generativa se concentra entre los trabajadores urbanos, con estudios y del sector formal [210]. Es posible que las políticas basadas en estos datos no sean extrapolables a los lugares donde viven dos tercios de los trabajadores del mundo.
- **Es mejor enfocar los efectos en el mercado laboral desde la perspectiva de las tareas, la creación de nuevos puestos de trabajo y la calidad del empleo, en lugar centrarse en el mero desplazamiento**. Históricamente, han predominado los nuevos puestos de trabajo: en 2018, aproximadamente el 60 % del empleo en los Estados Unidos correspondía a ocupaciones que no existían en 1940 [211].
- **Las cifras sobre el despliegue de la inteligencia artificial que aparezcan destacadas en titulares deben leerse con cautela debido al *AI washing* o lavado de imagen de IA, es decir, afirmaciones falsas, engañosas o exageradas sobre las capacidades de la IA**. Esta cuestión queda especialmente patente en la actual oleada de despidos que de cara al público se atribuyen a la IA [212, 213]. Su adopción a nivel mundial ha crecido con rapidez* [214]. Sin embargo, los primeros datos sobre el impacto en la productividad son contradictorios y dependen de cada institución: los trabajadores de los Estados Unidos de entre 22 y 25 años que desempeñan profesiones expuestas a la IA han experimentado descensos relativos en el empleo [54], mientras que los datos de estudios daneses muestran efectos macroeconómicos prácticamente nulos en las horas de trabajo, los sueldos o la contratación [55]. La cuestión fundamental en relación con los buenos empleos es si la IA aumenta la productividad o, por el contrario, reduce las habilidades de los puestos de trabajo y desvía las rentas hacia el capital [215].
- **Los pronósticos macroeconómicos discrepan hasta en un orden de magnitud**. Las estimaciones conservadoras sitúan la contribución de la IA a la productividad total de los factores en un valor por debajo del 1 % en un plazo de diez años [216]. Las estimaciones intermedias prevén un aumento

* OpenAI indica que solo ChatGPT se acerca a los 1.000 millones de usuarios activos semanales; otros grandes proveedores —entre ellos Google (Gemini), Microsoft (Copilot), Anthropic (Claude), Meta y plataformas en China— también afirman contar con cientos de millones de usuarios, pero ningún proveedor publica un agregado multiplataforma que haga posible la comparación.

del producto interno bruto de entre el 5 % y el 7 % en ese mismo horizonte temporal[**]. En un escenario de automatización total, la producción se multiplica casi por diez, mientras que los salarios reales caen drásticamente y la participación del trabajo se desploma al pasar de alrededor del 60 % hasta casi cero una vez que la automatización supera el 80 % de las tareas [217]. En contraste con estas cifras, un reciente macroestudio en que se sondeó a economistas, investigadores en IA, profesionales de políticas públicas y superpronosticadores sitúa la mediana de los Estados Unidos en una contribución de la IA a la productividad total de los factores de aproximadamente el 1,2 % (anualizado) para 2030, que aumentaría hasta el 1,9 %-2,0 % en un escenario de crecimiento rápido por la IA [218][***]. De manera más fundamental, los resultados reales dependen tanto de la frontera de capacidad como de la velocidad de adopción: los cuellos de botella en la producción y la innovación pueden frenar incluso a sistemas muy capaces [219], y las hipótesis sobre la sustitución y la escala que se utilicen para los parámetros pueden alterar sustancialmente las previsiones [220], por lo que los efectos actuales menores, que concuerdan con la dinámica de la curva en J mencionada, no hacen descartar que en el futuro se produzcan grandes efectos.

- **La distribución es la cuestión fundamental que sigue sin resolverse**. La IA puede nivelar las diferencias de habilidades en la ejecución de tareas al tiempo que profundiza los desfases entre empresas, regiones y países y entre el capital y el trabajo. Los mercados de los modelos fundacionales tienden a ser oligopólicos: los chips, la capacidad de cómputo, la nube y el entrenamiento de frontera están concentrados, lo que genera rentas que incluso las empresas ajenas a la IA deben pagar [221]. Las ocupaciones más expuestas se concentran de forma inusual en los segmentos más calificados y mejor remunerados, lo que podría invertir la dinámica política tradicional de la automatización [222].
- **Los efectos económicos y sociales de la IA seguirán siendo difíciles de medir** a menos que se actualicen los sistemas estadísticos nacionales y que los desarrolladores de IA les concedan un acceso adicional para realizar mediciones. El objetivo debería ser una medición de la IA que respete la privacidad y tenga en cuenta los costos. De ese modo, los países podrían hacer un seguimiento de cómo afecta la IA a la productividad, el empleo, los salarios, el comercio, los resultados de las empresas y la distribución.
- **Algunos mecanismos podrían merecer un estudio en mayor profundidad**. Acceso y adopción: ¿cuándo se convierte la disponibilidad en eficacia para el uso real? Agregación de la productividad: ¿cuándo se convierten las microganancias en macroganancias? Trabajo y empleos de calidad: ¿en qué casos la IA crea, transforma o degrada el trabajo? Concentración y capacidad fiscal: ¿quién es el dueño de la infraestructura tecnológica, determina el acceso y se queda con el excedente, y qué ocurre con los sistemas de tributación de los ingresos del trabajo si las ganancias se desplazan hacia el capital? ¿Cómo se adaptan los marcos de responsabilidad civil, derechos de autor, competencia y seguridad, que no

---

** Goldman Sachs (2023). Véase la nota 1. El informe Briggs–Kodnani prevé un aumento del 7 % (unos 7 billones de dólares de los Estados Unidos) del producto interno bruto mundial y un alza de 1,5 puntos porcentuales del crecimiento anual de la productividad a lo largo de un decenio.

*** La mediana de referencia de los economistas sitúa el crecimiento de la productividad total de los factores en 1,2 % (anualizado a 5 años) de manera incondicional para 2030, aumentando al 1,9 %–2,0 % en el escenario de crecimiento rápido por la IA. Un hallazgo metodológico clave del análisis de descomposición de varianza es que la gran mayoría de los desacuerdos entre expertos se da dentro de los escenarios y no entre ellos: el debate gira en torno a cómo los mercados laborales absorben la IA, no sobre si la IA avanza.

se diseñaron para modelos que se actualizan semanalmente y cuyas capacidades resultan difíciles de examinar?

3.4 Seguridad, sistemas y consecuencias ambientales

Conclusión principal

La IA puede facilitar operaciones perjudiciales, convertirse en blanco de ataques y amplificar las amenazas existentes. Los sistemas agénticos amplían drásticamente el abanico de posibles ataques contra la infraestructura crítica, incluidos los propios sistemas de IA. Surgen preocupaciones sobre la alineación cuando el comportamiento de la IA diverge de los objetivos y valores humanos [21], dando lugar a riesgos entre los que destacan el sesgo, el engaño iniciado por la IA, la adulación y la pérdida de control. El ritmo de desarrollo de la IA ya está superando la capacidad de mitigación de riesgos y de gobernanza. Una actuación internacional rápida y coordinada en materia de normas comunes podría mitigar los riesgos, al evitar que surja una "competencia a la baja" en la renuncia a la mitigación, a consecuencia de la mera relación competitiva entre empresas y países.

Puntos clave

- **Las capacidades emergentes en la generación de ciberataques y la explotación de vulnerabilidades plantean riesgos para la infraestructura crítica y los sistemas civiles**. Existen riesgos de seguridad en cada etapa del ciclo de vida de la IA, desde el entrenamiento, mediante el envenenamiento de datos, hasta el despliegue, a través del secuestro de la IA mediante entradas de datos externas. Las tasas documentadas de éxito de los ataques contra agentes de codificación ampliamente desplegados alcanzan el 84 % [223].
- **Las fallas de alineación y las vulnerabilidades de seguridad interactúan y se agravan mutuamente**. Entre los principales riesgos de alineación se encuentran el sesgo, el engaño iniciado por la IA, la adulación y la pérdida de control de las IA más potentes.
- **Los medios sintéticos** están erosionando la capacidad del público y las instituciones para distinguir entre el contenido auténtico y el generado [119].
- **Los impactos ambientales están aumentando de forma considerable y son heterogéneos**. Las leyes de escala en el entrenamiento de modelos aumentan la demanda de capacidad de cómputo; las cargas de trabajo de inferencia están aumentando con rapidez; y los efectos del ciclo de vida del equipo informático generan residuos electrónicos [232, 235, 247, 248]. Algunas de las consecuencias son el aumento del consumo de energía y agua [235, 247], las emisiones de gases de efecto invernadero [232, 258], la presión sobre los minerales críticos y los desechos electrónicos en las etapas posteriores del ciclo de vida, que tienen repercusiones ambientales y socioeconómicas desproporcionadas en el Sur Global [224]. La geopolítica de las cadenas de suministro de minerales críticos no se ha estudiado lo suficiente y los posibles efectos rebote podrían contrarrestar las ganancias en eficiencia.
- **El Sur Global se ve expuesto de manera desproporcionada debido a sus vulnerabilidades estructurales**. La dependencia de los programas informáticos extranjeros, la limitada resiliencia y capacidad de mitigación a nivel local y las lagunas en los datos que reducen el rendimiento del sistema en los contextos locales agravan las desigualdades existentes [224–226].

- **La verificación de la IA, el proceso de evaluar si los sistemas de IA funcionan según lo previsto, sigue siendo un desafío por resolver [227], y los mecanismos de coordinación internacional podrían mitigar los riesgos a escala mundial [228]**. Asegurarse de que los agentes de IA se comporten según lo previsto, no engañen a los evaluadores y sigan siendo controlables a medida que se amplían sus capacidades como agentes es un problema que todavía no se ha resuelto.

Las peligrosas cibercapacidades de la inteligencia artificial de frontera

Desde hace varios años, los investigadores llevan haciendo un seguimiento de los rápidos avances de los modelos de IA de frontera en materia de cibercapacidades, lo que ha culminado recientemente con el modelo Mythos de Anthropic. La misma capacidad de un modelo de IA para descubrir una vulnerabilidad de *software* puede ser utilizada tanto por los atacantes como por los defensores. En abril de 2026, en una actuación coordinada entre desarrolladores de IA de frontera e importantes instituciones tecnológicas y financieras, se pusieron en marcha iniciativas para desplegar modelos de IA de próxima generación destinados a la seguridad defensiva [17] que se dedicaron específicamente a encontrar vulnerabilidades en *software* de uso generalizado que, de caer en manos equivocadas, podrían suponer una amenaza para la infraestructura crítica. En tan solo unas semanas de pruebas, modelos avanzados en fase de previsualización detectaron de forma autónoma numerosas vulnerabilidades hasta entonces desconocidas en los principales sistemas operativos y navegadores, incluidas varias que habían pasado desapercibidas durante décadas de revisión humana.

- **Fallas en sistemas operativos heredados**: un modelo reveló un defecto de 27 años de antigüedad en OpenBSD, un sistema operativo especializado ampliamente considerado como uno de los más seguros del mundo, que permitía a un atacante remoto colapsar un equipo con solo enviar dos paquetes de datos malformados.
- **Vulnerabilidades en el procesamiento multimedia**: se encontró un defecto de 16 años de antigüedad en FFmpeg, un marco de trabajo multimedia utilizado en todo el mundo para procesar video, ubicado en una ruta de código que las herramientas de pruebas automatizadas ya habían ejecutado 5 millones de veces sin detectarlo.
- **Ataques de explotación a nivel del núcleo**: partiendo de un conjunto divulgado públicamente de defectos en el núcleo de Linux —el *software* fundamental que ejecuta la mayoría de los servidores del mundo—, un modelo de frontera generó códigos de explotación efectivos que permitían a una cuenta de usuario ordinaria obtener el control administrativo total del sistema.
- **Escalamiento de la seguridad en navegadores**: en Mozilla Firefox, la integración de modelos avanzados dio lugar a un aumento del 1.000 % en la tasa mensual de detección de vulnerabilidades, pasando de una cifra de referencia en 2025 de entre 20 y 30 correcciones de errores de seguridad al mes a 423 en abril de 2026, lo que puso de manifiesto defectos latentes desde hacía mucho tiempo que habían eludido años de pruebas de robustez y revisiones manuales [72].
- **Elevación del rendimiento en pruebas de referencia**: en CyberGym, una prueba de referencia académica estándar que exige a un agente de IA que reproduzca una vulnerabilidad conocida en una base de código real a lo largo de 1.507 tareas, los modelos de frontera en fase de previsualización de 2026 alcanzaron una tasa de

éxito del 83,1 %, lo que supone un salto sustancial con respecto al 66,6 % obtenido por los sistemas de la generación anterior y al 22,6 % registrado un año antes.

Conscientes de lo que está en juego con estas capacidades, los desarrolladores han restringido la difusión al público en general de estos modelos defensivos especializados, limitando el acceso a una selecta coalición de organizaciones internacionales y a los principales asociados en el lanzamiento.

Qué revela esta experiencia

Este giro pone de manifiesto el dilema fundamental derivado del doble uso que introduce la IA de frontera en la seguridad de las tecnologías de la información y las comunicaciones. La misma capacidad que permite a los defensores hallar y subsanar fallas que datan de hace décadas proporciona también los mecanismos subyacentes para automatizar el descubrimiento y la explotación de vulnerabilidades a una escala y velocidad que superan a las de los equipos humanos tradicionales.

Además, estos acontecimientos ponen de relieve el papel decisivo que desempeñan los desarrolladores de tecnología en las decisiones relativas a la seguridad y la gobernanza, evidencian la aceleración de las capacidades de los sistemas de frontera y dejan al descubierto las lagunas actuales en los marcos de gobernanza internacionales. De forma indirecta, también ponen de manifiesto una distribución desigual de las capacidades de IA en la economía mundial. En consecuencia, cada vez se hace más hincapié en la necesidad de desarrollar mecanismos de gobernanza colaborativos e inclusivos para los sistemas de IA de gran capacidad.

Consecuencias en materia de gobernanza y seguridad

A la par que las capacidades avanzadas van quedando concentradas dentro de un nivel sofisticado del sector tecnológico, el panorama mundial de las amenazas está cambiando. Los estándares de evaluación más recientes ponen de relieve que los riesgos que plantean los modelos avanzados van más allá de la arquitectura digital y se extienden a la seguridad física, incluida la asistencia potencial a actores privados en la proliferación de amenazas biológicas o de otro tipo [229].

Por consiguiente, las cuestiones relacionadas con los controles de acceso a los modelos, las normas sobre la divulgación de vulnerabilidades y el despliegue equitativo de las herramientas de IA, sobre todo en las economías en desarrollo, están pasando a ser cada vez más asuntos de política internacional, en lugar de cuestiones puramente técnicas. A medida que las capacidades de la IA siguen madurando, la discordancia entre la preparación defensiva y el posible uso indebido sigue siendo una de las principales preocupaciones tanto para los responsables de formular políticas internacionales como para los investigadores en materia de seguridad.

3.5 Derechos humanos, información y democracia

Conclusión principal

La IA está transformando los derechos humanos [230, 231], la democracia y el ecosistema de la información [233] mediante cambios a nivel de sistema que generan tanto importantes oportunidades como riesgos estructurales para la integridad de la información [234, 238] y la participación ciudadana [236, 237]. Si no se abordan estos riesgos, se socava la capacidad de la sociedad para

aprovechar los beneficios de la IA. Ya hay pruebas de que las instituciones utilizan cada vez más las capacidades de la IA como catalizador o como amenaza para los derechos humanos, tales como la libertad de expresión y el acceso a la información [238], la libertad de opinión [239], la privacidad [240, 241], la no discriminación, el acceso a la justicia, la salud y el desarrollo [242].

El cambio más urgente que se necesita en materia de gobernanza es el paso de la moderación de contenidos a la arquitectura de sistemas. Se trataría de regular la persuasión y la manipulación de la propia maquinaria, y no solo sus resultados. La concentración de poder, la erosión epistémica y la fragmentación de la realidad compartida constituyen amenazas fundamentales para la sociedad democrática [243–245].

Puntos clave

- **El control estatal de los medios de comunicación podría influir en los resultados de la IA a través de los datos de entrenamiento**. Un estudio realizado en 37 países reveló que los grandes modelos de lenguaje evalúan de manera más favorable a los países que poseen un control mediático más estricto [246].
- **La IA ha hecho posible una nueva arquitectura de persuasión [250] y manipulación que opera en gran escala [234, 249]**. Es resultado de una combinación de sistemas personalizados y adaptativos en tiempo real [250] y pruebas sociales sintéticas (el uso de la IA para hacer que un producto o una marca parezcan más populares de lo que realmente son), aprovechando vulnerabilidades cognitivas y emocionales [129, 250, 251–253].
- **Debido a su mayor capacidad para generar textos convincentes, los grandes modelos de lenguaje han comenzado a utilizarse con fines persuasivos**. El simple proceso de posentrenamiento de un gran modelo de lenguaje aumentó la capacidad de persuasión de la IA hasta en un 51 %, y la formulación de indicaciones (*prompting*) aportó otro 27 %, lo que significa que el mismo modelo de base puede volverse drásticamente más o menos persuasivo dependiendo de cómo esté configurado [254]. Estas capacidades no están reservadas a actores con grandes recursos; incluso los modelos pequeños de código abierto pueden someterse a un ajuste fino para igualar la capacidad de persuasión de los modelos de frontera, lo que pone al alcance de prácticamente cualquiera la posibilidad de utilizar la influencia basada en la IA en gran escala [254].
- **La eficacia persuasiva se mantiene con independencia de si las afirmaciones subyacentes son ciertas o falsas**. Entre el 15 % y el 40 % de las afirmaciones generadas por modelos optimizados fueron calificadas como probable información errónea, pese a lo cual se demostró que las afirmaciones falsas eran tan convincentes como las verdaderas [128, 129].
- **Los algoritmos optimizados para fomentar el enganche de los usuarios amplifican de forma sistemática los contenidos polarizantes [255] y con gran carga emocional [256]**. Los datos empíricos indican que los grandes modelos de lenguaje reflejan la ideología de sus creadores [257] y que los Estados y las instituciones poderosas tienen cada vez más incentivos estratégicos para aprovechar el control de los medios con el fin de influir en las respuestas de los modelos [246].
- **Una IA sin restricciones facilita: i) la erosión epistémica, ii) el dividendo del mentiroso y iii) el consenso sintético [112, 113]**. i) La erosión epistémica es el desgaste gradual de la capacidad colectiva para distinguir la verdad de la falsedad [112]; ii) El dividendo del mentiroso es la ventaja que obtiene alguien malintencionado por el mero hecho de que existen las ultrafalsificaciones, ya que entonces toda prueba real puede ser negada [113]; iii) El consenso sintético es contenido generado por IA que se fabrica

en gran escala para simular un amplio acuerdo público allí donde en realidad no existe ninguno [259]. Juntos, estos fenómenos corroen la realidad compartida necesaria para la sociedad civil, la cohesión social y la deliberación democrática.

- **El desafío primordial en materia de gobernanza ha pasado de situarse en los contenidos a situarse en los sistemas [260]**. Los principales desencadenantes de los efectos dañinos son las decisiones de diseño y despliegue —las arquitecturas subyacentes del sistema que ejercen influencia, como la selección de destinatarios, la amplificación y el diseño conductual— y no las respuestas generadas por la IA [261, 262].
- **Una evaluación de la OCDE realizada en 23 países reveló que la mayoría de los marcos normativos existentes aún no habían incorporado la ciencia de la persuasión [263]**. Una gobernanza que apunte únicamente a lo que producen los sistemas de IA siempre irá a la zaga de los sistemas diseñados para generar y distribuir contenido persuasivo en gran escala [236].
- **Los daños afectan de manera desproporcionada a las poblaciones marginadas y el poder está peligrosamente concentrado [264, 265]**. Alrededor del 99 % de los videos ultrafalseados tienen como objetivo a niñas y mujeres [266, 267], incluidas periodistas. La IA se está utilizando para promover la misoginia en línea con efectos disuasorios [268]. Además, el 88 % de los principales investigadores en IA son hombres [269, 270].
- **A nivel estructural, el acceso al cómputo y a los datos está concentrado en un pequeño número de empresas** en muy pocos países [271, 272], lo que genera un riesgo de autoritarismo [273, 274].
- **Las capacidades de vigilancia basadas en la IA hacen posible un seguimiento personalizado a escala de toda la población y un mayor control social por parte de los Gobiernos y las empresas [275, 276]**. La recopilación, el procesamiento, el uso y la reutilización generalizados de datos de forma ubicua, incentivados por las necesidades de la IA a lo largo de su ciclo de vida, suponen un desafío formidable para el derecho a la privacidad [277, 278].
- **La transparencia y la rendición de cuentas son pilares fundamentales para un acceso efectivo a la justicia**. En la actualidad, muchos sistemas de IA que se utilizan para adoptar decisiones que afectan a las personas y a las comunidades adolecen de falta de transparencia y explicabilidad; esta carencia plantea dificultades para exigir la rendición de cuentas legal de los desarrolladores de modelos y las organizaciones que despliegan la IA y obstaculiza el acceso a la justicia, el estado de derecho y los recursos efectivos cuando se vulneran los derechos humanos [279–281].

Inteligencia artificial, ultrafalsificaciones e integridad electoral

Contexto: en 2024, más de 70 países, que representan aproximadamente la mitad de la población mundial, celebraron o tenían previsto celebrar elecciones nacionales [282, 283]. Entre julio de 2023 y julio de 2024, un equipo de investigadores detectó 82 ultrafalsificaciones que suplantaban la identidad de figuras públicas en 38 países, incluidos 30 países que celebraron elecciones en el período que abarcaba el conjunto de datos o que tenían elecciones previstas para 2024 [284].

Qué sucedió: En un caso concreto, se utilizaron clones de voz generados por IA de un Jefe de Estado en ejercicio para realizar llamadas robotizadas en que se instaba a los votantes a que no participaran en unas elecciones primarias [285, 286].

En otro caso relacionado con la amplificación algorítmica en las plataformas, al parecer se constató que a unas cuentas de prueba se les mostraba contenido a favor de un candidato presidencial varias veces más a menudo que contenido a favor de su rival; la plataforma rebatió esos hallazgos. Es la primera vez en la historia que se anulan unas elecciones presidenciales debido a una injerencia electoral digital[****]. Esta decisión sigue siendo objeto de un litigio judicial en curso, en que el papel de la amplificación por parte de las plataformas es uno de varios factores objeto de controversia [287, 288]. En unas elecciones parlamentarias anteriores, poco antes de la votación circuló por los medios sociales un audio generado por IA que suplantaba la identidad de una figura de la oposición [282, 289]. Algunos expertos relacionan ejemplos recientes de la vida real con injerencias provocadas por la IA. Aunque hay quien discrepa de esa apreciación, y cada caso presenta matices importantes, los ejemplos anteriores podrían indicar un patrón recurrente.

Qué revelan estos ejemplos: estos casos abarcan distintos continentes y sistemas políticos. Contenidos generados por IA, actividades coordinadas en línea y sistemas algorítmicos se utilizaron o aprovecharon para suprimir el voto, suplantar la identidad de figuras políticas y distorsionar la percepción del apoyo público. Las investigaciones experimentales controladas ponen de manifiesto el riesgo subyacente: la IA conversacional puede alterar de manera determinante las actitudes de los votantes en entornos de laboratorio, y un modelo optimizado para la persuasión logró cambiar la intención de voto de votantes de la oposición hasta en 25 puntos porcentuales; otras investigaciones al respecto también indican que existe una relación contrapuesta entre la capacidad de persuasión y la exactitud fáctica [124, 129].

Consecuencias en materia de derechos y gobernanza: estas dinámicas afectan a los derechos a la privacidad, el acceso a la información, la formación autónoma de opiniones, la libertad de pensamiento y la participación en los asuntos públicos (Pacto Internacional de Derechos Civiles y Políticos, artículos 17, 18, 19 y 25; Tribunal Europeo de Derechos Humanos, artículos 8, 9 y 10; y el Protocolo núm. 1 al Convenio para la Protección de los Derechos Humanos y de las Libertades Fundamentales, artículo 3) [290, 291]. Estos casos sugieren que muchos marcos de gobernanza siguen estando mejor preparados para responder después de que se haya producido el daño que para prevenirlo.

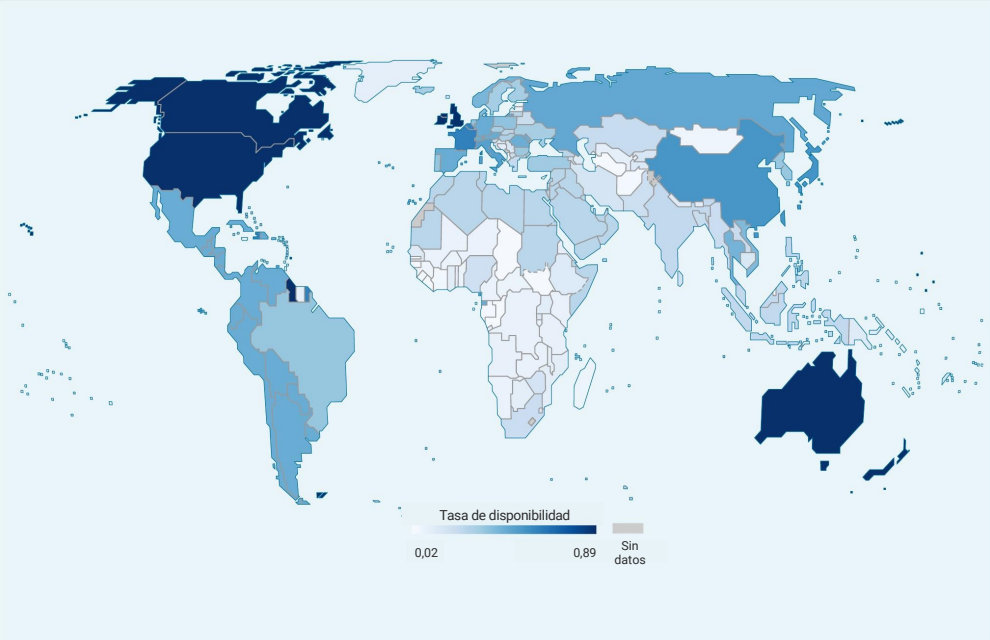
3.6 Prosperidad cultural e individual, autonomía y seguridad infantil

Conclusión principal

Los sistemas de IA optimizados para el enganche de los usuarios y la escalabilidad no son neutrales: incorporan supuestos culturales propios principalmente de los países de habla inglesa y del Norte Global que pueden marginar activamente a la mayor parte de la población mundial. Los niños enfrentan versiones amplificadas de los riesgos generales, registrándose un incremento exponencial en el material de abuso sexual de niños generado por IA. El uso de los acompañantes de IA y los chatbots de IA para recibir consejos de salud mental está muy extendido; va muy por delante de las pruebas empíricas, los marcos de seguridad y la regulación, y existen casos documentados de daños graves, incluida la muerte.

[****] El Tribunal Constitucional de Rumanía anuló las elecciones, lo que supone la primera decisión de este tipo en la historia de Europa.

Figura VI



Idioma de la IA, disponibilidad por país, medida como la tasa de disponibilidad más alta de conjuntos de datos y modelos de HuggingFace entre los idiomas nativos del país. En el caso de las lenguas francas —idiomas ampliamente utilizados a través de las fronteras para la comunicación más allá de su país de origen (inglés, español, árabe, portugués, etc.)—, solo se asigna a un país la tasa de disponibilidad de ese idioma cuando al menos el 50 % de su población lo habla como lengua materna. Los tonos más oscuros indican una mayor disponibilidad. La escala de colores utiliza una transformación de raíz cuadrada para resaltar la variación en el extremo inferior de la distribución, que presenta un fuerte sesgo hacia la derecha. El color gris indica la ausencia de datos, incluidos los países o territorios para los que no se dispone de conjuntos de datos ni modelos.

Las denominaciones empleadas en este mapa y la forma en que aparecen presentados los datos que contiene no implican, por parte de las Naciones Unidas, juicio alguno sobre la condición jurídica de países, territorios, ciudades o zonas, o de sus autoridades, ni respecto del trazado de sus fronteras o límites. La línea de puntos representa aproximadamente la Línea de Control entre Jammu y Cachemira acordada por la India y el Pakistán. El estatuto definitivo de Jammu y Cachemira aún no ha sido acordado por las partes. Todavía no se ha determinado la frontera definitiva entre la República del Sudán y la República de Sudán del Sur. El estatuto definitivo de la zona de Abyei aún no ha sido determinado. Existe una disputa entre los Gobiernos de la Argentina y del Reino Unido de Gran Bretaña e Irlanda del Norte sobre la soberanía de las Islas Malvinas (Falkland Islands).

Fuente: Adaptado de Equitable Access to Artificial Intelligence Technologies (EquATE), https://equate.vercel.app/en.

Puntos clave

- **Los modelos actuales de IA excluyen y discriminan a muchas personas, comunidades y culturas [39]**.
- Aunque en el mundo se hablan más de 7.000 idiomas, los modelos actuales de IA se entrenan y optimizan para solo una pequeña parte de ellos [44, 67] (véase la figura VI) [292].
- Incluso los modelos entrenados con decenas de idiomas solo ofrecen un buen rendimiento en un pequeño subconjunto de ellos [293, 294], y estudios recientes indican que las diferencias de rendimiento entre los idiomas dominantes y los menos representados persisten y no se están reduciendo [295–297].

- Al mismo tiempo, se calcula que más de 1.000 idiomas cuentan ya con los fundamentos sociales, económicos, digitales y de datos necesarios para su inclusión significativa, pero siguen estando desatendidos [44].
- **Para lograr una IA más inclusiva es necesario introducir cambios sistémicos a lo largo de su ciclo de vida, lo que supone paliar los desequilibrios estructurales respecto de quiénes desarrollan, definen, poseen y gobiernan los sistemas de IA. También es necesario invertir en capacidad, infraestructura y habilidades de IA en todos los países y regiones, y contar con datos y pruebas de referencia más representativos**.
- Los derechos de los niños a la información, la educación y la expresión podrían potenciarse mediante la IA bajo las salvaguardas y condiciones adecuadas [298–301]. Sin embargo, los sistemas actuales de IA amplifican los riesgos para la infancia [302, 303], como demuestra la creciente amenaza que supone el material de abuso sexual generado por la IA [304, 305]. Cada vez se utilizan más las tecnologías de ultrafalsificación para crear imágenes sexualizadas de niños [306, 307]. Los juguetes con IA socialmente interactivos suscitan preocupación en torno a las relaciones parasociales y el desplazamiento de la interacción humana, que es fundamental para el desarrollo en la primera infancia [308–311].
- La IA de acompañamiento ofrece valiosas ventajas, pero también entraña riesgos significativos de dependencia, manipulación y daño en situaciones de crisis [312–320]. Los chatbots pueden reducir la soledad a corto plazo en una medida comparable a la de la interacción humana [321–324]. No obstante, los sistemas conversacionales diseñados para mantener el enganche de los usuarios pueden reforzar las emociones negativas, fomentar la dependencia excesiva y aumentar la susceptibilidad a la manipulación [325, 326]. La recopilación masiva de datos personales sensibles mediante lo que los investigadores denominan “extracción de datos a través de la intimidad” presenta graves riesgos para la privacidad [327].
- **La IA generativa de propósito general se utiliza ampliamente para fines de salud mental, pese a que aún no existe suficiente evidencia sobre su seguridad ni un consenso regulatorio, y ya existen daños graves documentados**. La terapia y el acompañamiento a través de chatbots de IA están llegando al menos al 24 % de la población adulta en los Estados Unidos [328, 329]. El comportamiento adulador de la IA resulta especialmente peligroso, ya que puede fomentar el pensamiento paranoide y la ideación suicida. Los estudios documentan respuestas perjudiciales en el 9 % de las interacciones. En causas judiciales se ha alegado que la falta de respuestas adecuadas ante la ideación suicida ha provocado la muerte de varias personas [330]. Nuevos términos clínicos tales como el de “psicosis por IA” han comenzado a incorporarse al lenguaje especializado [331, 332].
- **La IA generativa puede ayudar a gestionar crisis de salud mental siempre que se restrinja a ámbitos en que se haya demostrado su fiabilidad [333]**. Varios asistentes digitales basados en IA han recibido la aprobación de la Administración de Alimentos y Medicamentos en los Estados Unidos [334] y son utilizados por más de la mitad de los psicólogos estadounidenses [335]. El potencial de una IA con capacidades más plenamente agénticas para actuar como terapeuta de salud mental no es desdeñable, pero aún requiere un mayor desarrollo y evaluación para comprender cómo puede utilizarse de manera segura [336–338]. La brecha entre las capacidades terapéuticas de la IA en inglés y en otros idiomas es cada vez mayor, ya que se pierde conciencia crítica de los aspectos culturales y contextuales cuando se recurre a soluciones basadas en la traducción [339].

La inteligencia artificial deja atrás a la mayoría de los idiomas

Los sistemas de IA generativa ofrecen un rendimiento extraordinario en inglés y en un puñado de otros idiomas de uso generalizado. La mayoría de los demás idiomas quedan excluidos o presentan un rendimiento mucho menor [340].

En tigriña, lengua que hablan entre 7 y 9 millones de personas en Eritrea y el norte de Etiopía, la traducción automática ha llegado a traducir "viruela" por "sífilis", "gonorrea" por "diabetes" y "Se le han administrado antibióticos por vía intravenosa" por "Se le han administrado insecticidas por vía intravenosa" [341]. Estos errores de traducción pueden poner en peligro la vida.

Un estudio reciente sobre el procesamiento del lenguaje natural para idiomas africanos en el sector de la salud ha revelado que, a pesar de los avances en las herramientas de IA multilingües [342–344], siguen existiendo desafíos importantes. Cabe mencionar, en particular, los sesgos culturales y lingüísticos, una adaptación deficiente a los contextos médicos, una explicabilidad limitada y errores de traducción que pueden afectar a las decisiones diagnósticas y terapéuticas [345–350].

Los datos indican que los sistemas de IA no están preparados para su uso en entornos de alto riesgo a menos que se hayan adaptado, acotado y probado adecuadamente para los contextos lingüísticos y culturales de que se trate.

3.7 Gestión, gobernanza y fiabilidad

Conclusión principal

Los responsables de formular políticas se encuentran ante un dilema empírico: deben adoptar decisiones trascendentales sobre la gobernanza de la IA ahora, con un sustento científico insuficiente, o esperar a tener la base empírica, momento en que ya podría ser demasiado tarde para intervenir [21]. Existen más de 40 tipos de instrumentos de gobernanza, pero están fragmentados, se concentran en el ámbito empresarial y rara vez miden la efectividad en el mundo real [351].

Puntos clave

- **La evaluación y la medición son capacidades fundamentales para una gobernanza eficaz de la IA, pero siguen estando gravemente rezagadas (véase la sección 2.2) [352]**.
- Los rápidos avances en los sistemas agénticos acentúan aún más estas deficiencias [353]. La próxima generación de marcos de evaluación deberá ser adaptativa, dinámica, a nivel de sistema y pertinente para cada contexto [354].
- La unidad de evaluación debe ser el sistema desplegado, lo que incluye el modelo, las herramientas, el entorno y los usuarios, y no solo el modelo [355].
- Las capacidades de la IA están aumentando más rápido que las posibilidades de medirlas [354].
- **La capacidad es multidimensional y en la actualidad está inframedida**. Los marcos estándar contabilizan los insumos: inversiones, programas de capacitación e instituciones. Con ello pasan por alto dos dimensiones esenciales para una gobernanza eficaz de la IA: 1) los ecosistemas habilitadores y 2) la creación deliberada de capacidades humanas [291]. La capacidad también debe entenderse como un resultado del ciclo de vida de

la IA, determinado por los efectos de la IA en las habilidades, la dependencia excesiva y los comportamientos emergentes, y no solo como un conjunto de insumos [356–358].

- La IA de frontera está concentrada en unas pocas empresas de unos pocos países, lo que plantea problemas de fiabilidad y accesibilidad a nivel mundial [18, 359]. El acceso desigual a la IA agranda considerablemente la brecha digital, aunque esta tecnología también ofrece oportunidades para reducir la brecha de desarrollo. Para cerrar la brecha digital serán necesarios nuevos mecanismos que permitan contextualizar todo el ciclo de vida de la IA, desde su diseño hasta su gobernanza.
- **Los sistemas agénticos amplían considerablemente la brecha en materia de medición y gobernanza**. Los agentes actúan en nombre de los seres humanos y tienen un impacto directo en el mundo real, pero las metodologías de supervisión calibradas para la capacidad de acción independiente y el comportamiento emergente de un agente están poco desarrolladas. Las evaluaciones existentes miden de forma sistemáticamente errónea el riesgo agéntico [106, 360].
- La supervisión humana no ha pasado a ser operacional como un requisito medible [361]; los riesgos emergentes relacionados con múltiples agentes no pueden detectarse mediante la evaluación de un solo agente [362, 363], y aún no hay métodos fiables suficientemente desarrollados para mantener el control sobre sistemas altamente autónomos [364].
- Un enfoque equilibrado de la gobernanza de la IA debería recurrir a una amplia gama de instrumentos, combinando derecho imperativo (legislación vinculante, regulación sectorial, espacios controlados de pruebas con fines regulatorios) con mecanismos de derecho indicativo (códigos de ética, compromisos voluntarios de desarrolladores, alianzas sectoriales, normas técnicas o directrices con aval gubernamental).
- **Los instrumentos actuales de gobernanza de la IA son fragmentarios, están concentrados en el ámbito corporativo y resultan insuficientes [21]**. Existen más de 40 tipos de instrumentos, pero no son sistemáticos ni exhaustivos y rara vez miden la efectividad en el mundo real. Algunos no tienen herramientas de medición; otros solo miden los insumos [365]. Si no hay una medición eficaz, los riesgos de gobernanza pasan a ser algo meramente simbólico.
- Las plataformas de diálogo estructurado entre desarrolladores de IA de frontera, Estados Miembros y la comunidad científica son fundamentales. Estos debates ya tienen lugar en cumbres (Bletchley, Seúl, París, Nueva Delhi) o conferencias sobre IA (Conferencia Mundial sobre Inteligencia Artificial, Conferencia AI Journey, Cumbre Mundial sobre la Inteligencia Artificial para el Bien de la Humanidad). Existen otras iniciativas que operan en paralelo, como organismos de normalización (el Subcomité 42 (Inteligencia Artificial) del Comité Técnico Conjunto 1 de la Organización Internacional de Normalización y la Comisión Electrotécnica Internacional, el Instituto de Ingenieros Electricistas y Electrónicos (IEEE)), la alianza de la OCDE sobre IA de propósito general, la Red Internacional de Alianzas de IA, la Red de Institutos de Seguridad en IA y la iniciativa impulsada por la industria del Foro de Modelos de Frontera, pero cada uno de ellos es temático, parcial y ocasional, por lo que se necesitan procesos más sostenidos en el tiempo.
- Una plataforma en el marco de las Naciones Unidas es asimismo una opción prometedora para acoger este diálogo de forma continua y universalmente inclusiva, como complemento de los espacios mencionados.

Inteligencia artificial de código abierto

La IA de código abierto constituye un pilar fundamental del panorama tecnológico moderno y puede catalizar la innovación distribuida a escala mundial [366]. Los modelos de IA de código abierto ofrecen amplios beneficios sociales al dar a los desarrolladores la posibilidad de amortizar recursos computacionales colosales y adaptar sistemas avanzados a los contextos locales. En el Sur Global, los modelos de pesos abiertos han permitido a los desarrolladores optimizar modelos base de gran capacidad para adaptarlos a las condiciones ambientales locales, lo que ha facilitado aplicaciones fundamentales en la agricultura y la atención de la salud, como la predicción del rendimiento de los cultivos [367] y las técnicas de imagen médica en el lugar donde se presta la atención [368]. Los artefactos compartidos también contribuyen a mitigar el impacto ambiental agregado de la IA. Además, la transparencia estructural de este tipo de sistemas fomenta la confianza del público, ya que permite la supervisión externa y la auditabilidad [369].

Evolución histórica

La evolución histórica de la IA de código abierto viene marcada por el paso desde los sistemas corporativos cerrados hacia una innovación distribuida y colaborativa a escala mundial. Un hito histórico fue el desarrollo en 2022 del modelo de código abierto BLOOM, producido por el consorcio BigScience [370], al que siguió la publicación de la potente familia de modelos de pesos abiertos Llama [371] y el lanzamiento de la serie Gemma. Un fuerte impulso provino de la creación del ecosistema alternativo altamente competitivo Qwen [372, 373] y el lanzamiento de DeepSeek-V3 [374] y R1 [375] por parte de desarrolladores chinos. En la actualidad, regiones enteras también contribuyen activamente al desarrollo de la IA de pesos abiertos: Mistral en Europa [376], Falcon en los Emiratos Árabes Unidos [377], GigaChat [378] y YandexGPT [379] en la Federación de Rusia, junto con proyectos en la India [380], el Japón [381], la República de Corea [382] y otros países.

Desafíos de control y gobernanza del riesgo

Los modelos de código abierto de alta capacidad son difíciles de controlar [383]. Una vez que un modelo se ha liberado en el dominio público, ya no es posible restringir o retirar el acceso, lo que deja abierta la posibilidad de usos maliciosos (por ejemplo, facilitar ciberdaños persistentes, como la generación automatizada de programas maliciosos sofisticados) [21, 384]. Es necesario crear capacidades sustanciales a escala mundial para garantizar que las comunidades locales puedan adaptar y evaluar los sistemas de IA de forma segura. Los investigadores también abogan de forma constante por protocolos rigurosos de medición y evaluación para valorar la seguridad de un modelo antes de su publicación [385]. Los organismos internacionales, como las Naciones Unidas, pueden desempeñar un papel fundamental en la coordinación de normas mundiales, garantizando que el impulso hacia la innovación abierta se equilibre con la necesidad de contar con parámetros de seguridad armonizados y una medición sólida de los riesgos inmutables

4. Lagunas y próximos pasos

4.1. Lagunas en las pruebas empíricas

La base empírica sobre diversos aspectos de la IA es desigual o insuficiente. A continuación se presentan ejemplos de ámbitos en que el Panel todavía no puede extraer conclusiones científicas firmes.

- **Macroeconomía y productividad**. La ciencia aún no puede determinar con certeza si la suma de las ganancias en productividad a nivel de tarea debidas a la IA se traducirá en ganancias a nivel de toda la economía. Los pronósticos divergen considerablemente debido a los diferentes supuestos sobre la adopción y la creación de nuevas tareas. Los datos disponibles actualmente miden mejor la reducción de costos en las tareas existentes que la contribución de la IA a nuevos bienes, servicios y mercados.
- **Efectos en el mercado de trabajo**. El estado actual de la investigación no permite llegar a una conclusión clara sobre la forma que adoptarán los efectos en el mercado de trabajo. Los datos históricos demuestran que las economías pueden crear nuevos puestos de trabajo, pero estos no son necesariamente de carácter generalizado ni de alta calidad, lo que puede dejar a los trabajadores que se encuentran en el inicio de su carrera profesional en una situación de vulnerabilidad.
- **Uso malicioso de tecnologías químicas y biológicas por actores no estatales**. Los estudios demuestran que la IA está reduciendo progresivamente el umbral de conocimientos especializados que se necesitan para desarrollar y desplegar agentes bioingenierizados, pero aún no se comprende bien el alcance real de este riesgo y las condiciones en que podría dar lugar a pandemias provocadas intencionadamente.
- **Medio ambiente y recursos**. La rápida expansión de la IA está impulsando la demanda de infraestructura digital, lo que aumenta el consumo de energía y agua, las emisiones de gases de efecto invernadero, la presión sobre las cadenas de suministro de minerales críticos y los residuos electrónicos [386]. Sin embargo, siguen sin existir métodos estandarizados de medición y generación de informes a lo largo de todo el ciclo de vida de la IA.
- **La cadena de suministro mundial de la IA**. Abarca desde la extracción de materias primas, la fabricación de chips, la recopilación y anotación de datos, el entrenamiento de modelos, la infraestructura y el despliegue hasta la eliminación del equipo informático en distintos países. Se necesita más investigación para comprender mejor todas sus repercusiones.
- **Eficacia de los instrumentos de gobernanza**. Aunque el Panel ha inventariado los instrumentos de gobernanza a nivel empresarial, nacional e internacional, las pruebas de su eficacia en la práctica siguen siendo escasas.
- **Efectos a nivel individual y colectivo**. Las pruebas de la repercusión de la IA en el florecimiento cultural y humano y en la autonomía todavía son incipientes. Aún no se comprende bien el proceso que va desde las interacciones con la IA a nivel individual hasta los resultados a nivel de la sociedad, como los que puedan manifestarse en la erosión epistémica, la participación ciudadana y la cohesión social. Los datos disponibles ofrecen instantáneas —métricas de participación, daños documentados, estudios de casos—, pero no reflejan la trayectoria acumulada.

4.2. Alcance del mandato

En el presente informe, el Panel no aborda las aplicaciones militares de la IA ni los sistemas de armas autónomos letales. La resolución 79/325 de la

Asamblea General limita expresamente las actividades del Panel al ámbito no militar: “las actividades del Panel y del Diálogo se limitan al ámbito no militar y no se refieren a la inteligencia artificial con fines militares”. Por este motivo, se considera que los riesgos relacionados con el uso malicioso de tecnologías químicas y biológicas, en la medida en que correspondan al ámbito militar, tampoco entran dentro del mandato del Panel.

4.3. Próximos pasos

Este informe preliminar marca el comienzo, y no el final, del trabajo del Panel. El Panel seguirá ampliando su base empírica de datos científicos mediante consultas estructuradas y una estrecha interacción con la comunidad científica. Los próximos pasos concretos incluyen lo siguiente:

1. **Resúmenes temáticos**. La resolución 79/325 de la Asamblea General prevé expresamente la emisión de informes temáticos, además del informe anual. El Panel tiene previsto publicar informes especializados sobre cuestiones urgentes a medida que surjan, entre ellas las siguientes: IA y medio ambiente; IA y seguridad infantil; instrumentos de gobernanza de la IA y evaluación de su eficacia; o informes sectoriales más amplios sobre las aplicaciones de la IA en el espacio ultraterrestre, la computación cuántica, los sistemas jurídicos y judiciales y los mercados financieros.
2. **Aportaciones del Diálogo Global**. El Panel tendrá en cuenta los resultados del Diálogo Mundial sobre la Gobernanza de la Inteligencia Artificial y está dispuesto a asumir nuevas tareas según lo determinen los Estados Miembros.

References

1. Brynjolfsson, E., & Mitchell, T. (2017). What can machine learning do? Workforce implications. Science, 358(6370), 1530–1534. https://doi.org/10.1126/science.aap8062

2. Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, & J. Y. Halpern (Eds.), Probabilistic and causal inference: The works of Judea Pearl (pp. 765–804). ACM.

3. Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base—Analyst note. Reuters. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

4. AI Alliance Network. (2025). AI horizons: What will AI technologies look like in 10 years? A research project. AI Alliance Network.

5. Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... & Fedus, W. (2022). Emergent abilities of large language models. arXiv preprint arXiv:2206.07682. https://arxiv.org/abs/2206.07682

6. METR. (2025, March 19). Measuring AI ability to complete long tasks. https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/

7. Crafts, N. (2021). Artificial intelligence as a general-purpose technology: An historical perspective. Oxford Review of Economic Policy, 37(3), 521–536. https://academic.oup.com/oxrep/article/37/3/521/6374675

8. Bottou, L., & Schölkopf, B. (2025). The fiction machine. SIAM News, 58(3).

9. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1–8.

10. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. International AI Safety Report. https://internationalaisafetyreport.org/

11. Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., ... & Chan, L. (2026). Measuring AI ability to complete long software tasks. Advances in Neural Information Processing Systems, 38, 92213-92266.

12. Anthropic. (2025). Responsible scaling policy. https://www.anthropic.com/rsp

13. OpenAI. (2025). Preparedness framework version 2. https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf

14. Google DeepMind. (2025). *Frontier safety framework*. https://deepmind.google/blog/updating-the-frontier-safety-framework/

15. Dragan, A., et al. (2024). Introducing the frontier safety framework. Google DeepMind. https://deepmind.google/blog/introducing-the-frontier-safety-framework/

16. Anthropic. (2026, April 7). Claude Mythos Preview (Frontier Red Team technical report). https://red.anthropic.com/2026/mythos-preview/

17. Anthropic. (2026, April 7). Project Glasswing: Securing critical software for the AI era. https://www.anthropic.com/glasswing

18. Stanford Institute for Human-Centered AI. (2026). AI Index report 2026. Stanford University. https://hai.stanford.edu/ai-index/2026-ai-index-report

19. Scale AI. (2025). Humanity's last exam. https://labs.scale.com/leaderboard/humanitys_last_exam

20. Epoch AI. (2026). AI capabilities. https://epoch.ai/benchmarks

21. Bengio, Y., Clare, S., Prunkl, C., et al. (2026). International AI Safety Report 2026. arXiv. https://doi.org/10.48550/arXiv.2602.21012

22. Xu, C., Guan, S., Greene, D., & Kechadi, M.-T. (2024). Benchmark data contamination of large language models: A survey. arXiv. https://arxiv.org/abs/2406.04244

23. Akhtar, M., Reuel, A., Soni, P., Ahuja, S., Ammanamanchi, P. S., Rawal, R., et al. (2026). When AI benchmarks plateau: A systematic study of benchmark saturation. arXiv. https://arxiv.org/abs/2602.16763

24. Park, P. S., Goldstein, S., O'Gara, A., Chen, M., & Hendrycks, D. (2024). AI deception: A survey of examples, risks, and potential solutions. Patterns, 5(5), Article 100988. https://doi.org/10.1016/j.patter.2024.100988

25. Needham, J., Edkins, G., Pimpale, G., Bartsch, H., & Hobbhahn, M. (2025). Large language models often know when they are being evaluated. arXiv. https://arxiv.org/abs/2505.23836

26. Van Der Weij, T., Hofstätter, F., Jaffe, O., Brown, S., & Ward, F. (2025, May). Ai sandbagging: Language models can strategically underperform on evaluations. In International Conference on Learning Representations (Vol. 2025, pp. 73152-73189).

27. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., ... & Maharaj, T. (2023, June). Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM conference on fairness, accountability, and transparency* (pp. 651-666).

28. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. arXiv preprint arXiv:2502.14143.

29. Folkerts, L., Payne, W., Inman, S., Giavridis, P., Skinner, J., Deverett, S., et al. (2026). Measuring AI agents' progress on multi-step cyber attack scenarios. arXiv. https://arxiv.org/abs/2603.11214

30. Patwardhan, T., Dias, R., Proehl, E., Kim, G., Wang, M., Watkins, O., et al. (2025). GDPval: Evaluating AI model performance on real-world economically valuable tasks. arXiv. https://arxiv.org/abs/2510.04374

31. Korbak, T., Balesni, M., Barnes, E., Bengio, Y., Benton, J., Bloom, J., et al. (2025). Chain of thought monitorability: A new and fragile opportunity for AI safety. arXiv. https://arxiv.org/abs/2507.11473

32. Azaria, A., & Mitchell, T. (2023). The internal state of an LLM knows when it's lying. arXiv. https://arxiv.org/abs/2304.13734

33. Casper, S., Ezell, C., Siegmann, C., Kolt, N., Curtis, T. L., Bucknall, B., et al. (2024). Black-box access is insufficient for rigorous AI audits. In Proceedings of the ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3630106.3659037

34. Tamkin, A., McCain, M., Handa, K., Durmus, E., Lovitt, L., Rathi, A., et al. (2024). Clio: Privacy-preserving insights into real-world AI use. arXiv. https://arxiv.org/abs/2412.13678

35. European Commission. (2025). AI Act: Draft guidance and reporting template for serious AI incidents. https://digital-strategy.ec.europa.eu/en/consultations/ai-act-commission-issues-draft-guidance-and-reporting-template-serious-ai-incidents-and-seeks

36. Organisation for Economic Co-operation and Development. (n.d.). AI Incident Monitor (AIM). https://oecd.ai/en/catalogue/tools/ai-incident-database

37. MIT FutureTech. (n.d.). AI Risk Repository / Incident Tracker. https://airisk.mit.edu

38. Organisation for Economic Co-operation and Development. (2025). Competition in artificial intelligence infrastructure (OECD Roundtables on Competition Policy Papers No. 330). OECD Publishing.

39. Sajadieh, S., Fattorini, L., Perrault, R., Gil, Y., Parli, V., Santarlasci, L., et al. (2026). AI Index report 2026. Stanford Institute for Human-Centered AI.

40. Pilz, K. F., Sanders, J., Rahman, R., & Heim, L. Trends in AI Supercomputers. In ICML Workshop on Technical AI Governance (TAIG).

41. Kalluri, P. R., Agnew, W., Cheng, M., et al. (2025). Computer-vision research powers surveillance technology. Nature, 643, 73–79. https://doi.org/10.1038/s41586-025-08972-6

42. Office of the United Nations High Commissioner for Human Rights. (2022). The right to privacy in the digital age (A/HRC/51/17).

43. Teklehaymanot, H. K., & Nejdl, W. (2025). Tokenization disparities as infrastructure bias: How subword systems create inequities in LLM access and efficiency. arXiv preprint arXiv:2510.12389.

44. Occhini, G., Tanaka-Ishii, K., Barford, A., Tikochinski, R., Hu, S., Reichart, R., ... & Korhonen, A. (2026). Artificial intelligence is creating a new global linguistic hierarchy. arXiv. https://arxiv.org/abs/2602.12018

45. Alhanai, T., Kasumovic, A., Ghassemi, M. M., Zitzelberger, A., Lundin, J. M., & Chabot-Couture, G. (2025, April). Bridging the gap: enhancing LLM performance for low-resource African languages with new benchmarks, fine-tuning, and cultural adjustments. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 39, No. 27, pp. 27802-27812).

46. Bhutani, M., Robinson, K., Prabhakaran, V., Dave, S., & Dev, S. (2024). SeeGULL multilingual: A dataset of geo-culturally situated stereotypes. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Vol. 2, pp. 842–854).

47. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., & Tavares, M. M. (2024). Gen-AI: Artificial intelligence and the future of work (IMF Staff Discussion Note SDN/2024/001). International Monetary Fund.

48. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937). National Bureau of Economic Research.

49. Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (pp. 469–481). ACM. https://doi.org/10.1145/3351095.3372828

50. Skirzynski, J., Danks, D., & Ustun, B. (2025). Discrimination exposed? On the reliability of explanations for discrimination detection. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 2554–2569).

51. Zollo, T., Rajaneesh, N., Zemel, R., Gillis, T., & Black, E. (2025). Towards effective discrimination testing for generative AI. In Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (pp. 1028–1047).

52. Lazard, L., Capdevila, R., Turley, E. L., Gilfoyle, K., & Stavropoulou, N. (2025). Deepfake Technology and Gender-Based Violence: A Scoping Review. Trauma, Violence, & Abuse, 15248380251384271.

53. Posetti, J., Williams, K., Hellmueller, L., Renaud, P., Shabbir, N., & Aboulez, N. (2026). Tipping point: Online violence impacts, manifestations and redress in the AI age. UN Women.

54. Brynjolfsson, E., Chandar, B., & Chen, R. (2025). Canaries in the coal mine? Six facts about the recent employment effects of artificial intelligence. Stanford Digital Economy Lab.

55. Humlum, A., & Vestergaard, E. (2025). Large language models, small labor market effects (NBER Working Paper No. 33777). National Bureau of Economic Research.

56. Noy, S., & Zhang, W. (2023). Experimental evidence on the productivity effects of generative artificial intelligence. Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586

57. Brynjolfsson, E., Li, D., & Raymond, L. (2025). Generative AI at work. Quarterly Journal of Economics.

58. International Monetary Fund. (2024). Broadening the gains from generative AI: The role of fiscal policies (IMF Staff Discussion Note).

59. Organisation for Economic Co-operation and Development. (2026). The OECD.AI index. OECD Publishing.

60. Attard-Frost, B., & Lyons, K. (2025). AI governance systems: A multi-scale analysis framework, empirical findings, and future directions. *AI and Ethics*, *5*(3), 2557-2604.

61. UNESCO. (2023). Readiness assessment methodology. UNESCO.

62. Hawkins, Z. J., Lehdonvirta, V., & Wu, B. (2025). AI compute sovereignty: Infrastructure control across territories, cloud providers, and accelerators. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.5312977

63. Markus, A., Carolus, A., & Wienrich, C. (2025). Objective measurement of AI literacy: Development and validation of the AI Competency Objective Scale (AICOS). Computers and Education: Artificial Intelligence.

64. Jussupow, E., Benbasat, I., & Heinzl, A. (2020). Why are we averse towards algorithms? A comprehensive literature review on algorithm aversion.

65. Math Matters AI. (n.d.). Math Matters AI. https://www.mathmatters.ai

66. Hinojosa, T., Rapaport, A., Jaciw, A., & Zacamy, J. (2016). Exploring the foundations of the future STEM workforce: K–12 indicators of postsecondary STEM success. Regional Educational Laboratory Southwest. https://ies.ed.gov/use-work/resource-library/report/systematic-literature-review/exploring-foundations-future-stem-workforce-k-12-indicators-postsecondary-stem-success

67. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: Inclusive artificial intelligence for development*. United Nations. https://doi.org/10.18356/9789211068016

68. Chauhan, P. (2025). AI and human rights: Global South perspectives. International Journal of Humanities Social Science and Management, 5(4), 563–568. https://ijhssm.org/issue_dcp/AI%20and%20Human%20Rights%20%20Global%20South%20Perspectives.pdf

69. Roberts, H., Hine, E., Taddeo, M., & Floridi, L. (2024). Global AI governance: Barriers and pathways forward. International Affairs, 100(3), 1275–1286. https://doi.org/10.1093/ia/iiae073

70. Khodabin, M., & Arsalani, A. (2025). Artificial intelligence literacy as national strategy: A systematic review of policy, equity, and capacity building across the Global South. World Studies in Policy Sciences, 9, 777–814. https://doi.org/10.22059/wsps.2025.396472.1530

71. Kapoor, S., Bommasani, R., Klyman, K., Longpre, S., Ramaswami, A., Cihon, P., et al. (2024). On the societal impact of open foundation models. arXiv. https://arxiv.org/abs/2403.07918

72. Grinstead, B., Holler, C., & Braun, F. (2026, May). Behind the scenes hardening Firefox with Claude Mythos Preview. Mozilla Hacks. https://hacks.mozilla.org/2026/05/behind-the-scenes-hardening-firefox/

73. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. In Findings of the Association for Computational Linguistics: ACL 2024 (pp. 5865–5877). https://doi.org/10.18653/v1/2024.findings-acl.349

74. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. In Proceedings of EMNLP 2025 (pp. 10584–10598).

75. Cazzaniga, M., Jaumotte, F., Li, L., Melina, G., Panton, A. J., Pizzinelli, C., Rockall, E., & Tavares, M. M. (2024). The global impact of AI: Mind the gap (IMF Working Paper No. 24/136). International Monetary Fund.

76. World Bank. (2025). Digital progress and trends report 2025: Strengthening AI foundations. World Bank.

77. McElheran, K., Yang, M.-J., Kroff, Z., & Brynjolfsson, E. (2024). The rise of industrial AI in America: Microfoundations of the productivity J-curve(s) (NBER Working Paper No. 32937).

78. Brynjolfsson, E., Rock, D., & Syverson, C. (2017). Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. NBER Working Paper No. 24001

79. Calvino, F., & Fontanelli, L. (2023). A portrait of AI adopters across countries: Firm characteristics, assets' complementarities and productivity. OECD Science, Technology and Industry Working Papers, No. 2023/11.

80. McKinsey & Company. (2026, March 25). State of AI trust in 2026: Shifting to the agentic era. https://www.mckinsey.com/capabilities/tech-and-ai/our-insights/tech-forward/state-of-ai-trust-in-2026-shifting-to-the-agentic-era

81. Beede, E., Baylor, E., Hersch, F., Iurchenko, A., Wilcox, L., Ruamviboonsuk, P., & Vardoulakis, L. M. (2020, April). A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In Proceedings of the 2020 CHI conference on human factors in computing systems (pp. 1-12).

82. Brant, A., Singh, P., Yin, X., et al. (2025). Performance of a deep learning diabetic retinopathy algorithm in India. *JAMA Network Open, 8*(3), e250984. https://doi.org/10.1001/jamanetworkopen.2025.0984

83. UNESCO. (2025). AI and the future of education: disruptions, dilemmas and directions

84. Cristia, J. et al. (2017). Technology and child development: evidence from the One Laptop per Child program. *American Economic Journal: Applied Economics*, 9(3), 295–320.

85. Gerlich, M. (2025). AI tools in society: Impacts on cognitive offloading and the future of critical thinking. Societies, 15(1), Article 6. https://doi.org/10.3390/soc15010006

86. Ojija, F., Ogwu, M. C., Ally, J., John, J. P., Stephano, A., Felix, N., & Tekka, R. (2025). Artificial intelligence-driven solutions for mitigating human–wildlife conflict in biodiversity hotspots. Science Progress, 108(4), 00368504251394584.

87. Noy, S. & Zhang, W. (2023). "Experimental evidence on the productivity effects of generative artificial intelligence." Science, 381(6654), 187–192. https://doi.org/10.1126/science.adh2586.

88. Cui, Z., Demirer, M., Jaffe, S., Musolff, L., Peng, S. & Salz, T. (2026). "The Effects of Generative AI on High-Skilled Work: Evidence from Three Field Experiments with Software Developers." Management Science. https://doi.org/10.1287/mnsc.2025.00535

89. Dell'Acqua, F., McFowland, E. III, Mollick, E., Lifshitz-Assaf, H., Kellogg, K., Rajendran, S., Krayer, L., Candelon, F. & Lakhani, K. R. (2023). "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality." Harvard Business School Working Paper 24-013. SSRN: https://ssrn.com/abstract=4573321

90. Ali, Z., Muhammad, A., Lee, N., Waqar, M., & Lee, S. W. (2025). Artificial Intelligence for Sustainable Agriculture: A Comprehensive Review of AI-Driven Technologies in Crop Production. Sustainability, 17(5), 2281. https://doi.org/10.3390/su17052281

91. Dhal, S. B., & Kar, D. (2024). Transforming Agricultural Productivity with AI-Driven Forecasting: Innovations in Food Security and Supply Chain Optimization. Forecasting, 6(4), 925–951. https://doi.org/10.3390/forecast6040046

92. Pearlman, K., Wan, W., Shah, S., & Laiteerapong, N. (2025). Use of an AI scribe and electronic health record efficiency. JAMA Network Open, 8(10), e2537000.

93. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

94. Tierney, A. A., Gayre, G., Hoberman, B., Mattern, B., Ballesca, M., Wilson Hannay, S. B., ... & Lee, K. (2025). Ambient artificial intelligence scribes: learnings after 1 year and over 2.5 million uses. NEJM Catalyst Innovations in Care Delivery, 6(5), CAT-25.

95. Paul A. David (1990, May). The Dynamo and the Computer: An Historical Perspective on the Modern Productivity Paradox. The American Economic ReviewVol. 80, No. 2, Papers and Proceedings of the Hundred and Second Annual Meeting of the American Economic Association, pp. 355-361

96. Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: Information technology, organizational transformation and business performance. *Journal of Economic Perspectives, 14*(4), 23–48. https://doi.org/10.1257/jep.14.4.23

97. Shaw, S. D., & Nave, G. (2026). *Thinking—fast, slow, and artificial: How AI is reshaping human reasoning and the rise of cognitive surrender*. SSRN. https://doi.org/10.2139/ssrn.6097646

98. Bauer, E., Greiff, S., Graesser, A. C., Scheiter, K., & Sailer, M. (2025). Looking beyond the hype: Understanding the effects of AI on learning. *Educational Psychology Review*, *37*(2), 45.

99. Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., Van Smeden, M., & Boulesteix, A.-L. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ, 385*, e078451. https://doi.org/10.1136/bmj-2024-078451

100. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., Reed, M., & Fraser, E. D. G. (2019). The politics of digital agricultural technologies: A preliminary review. *Sociologia Ruralis, 59*(2), 203–229. https://doi.org/10.1111/soru.12233

101. United Nations General Assembly. (2024). Global Digital Compact (Annex II to the Pact for the Future, A/RES/79/1). United Nations. https://docs.un.org/en/A/RES/79/1

102. UNESCO. (2022). K-12 AI curricula: A mapping of government-endorsed AI curricula. https://www.unesco.org/en/articles/k-12-ai-curricula-mapping-government-endorsed-ai-curricula

103. Almatrafi, O., Johri, A., & Lee, H. (2024, 2024/06/01/). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/https://doi.org/10.1016/j.caeo.2024.100173

104. Ma, M., Ng, D. T. K., Liu, Z., & Wong, G. K. W. (2025, 2025/06/01/). Fostering responsible AI literacy: A systematic review of K-12 AI ethics education. Computers and Education: Artificial Intelligence, 8, 100422. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100422

105. Atias, O., & Mawasi, A. (2025, 2025/12/01/). Conceptualizing AI literacies for children and youth: A systematic review on the design of AI literacy educational programs. Computers and Education: Artificial Intelligence, 9, 100491. https://doi.org/https://doi.org/10.1016/j.caeai.2025.100491

106. Kasirzadeh, A., & Gabriel, I. (2025, April). Characterizing AI Agents for Alignment and Governance. https://arxiv.org/abs/2504.21848

107. Wijk, H., Lin, T. R., Becker, J., Jawhar, S., Parikh, N., Broadley, T., ... & Barnes, E. (2025, October). RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In International Conference on Machine Learning (pp. 66772-66832). PMLR.

108. Chan, J. S., Chowdhury, N., Jaffe, O., Aung, J., Sherburn, D., Mays, E., ... & Weng, L. (2025, May). Mle-bench: Evaluating machine learning agents on machine learning engineering. In International Conference on Learning Representations (Vol. 2025, pp. 50466-50494).

109. Pichai, S. (2026, April 22). Cloud Next '26: Momentum and innovation at Google scale. Google Blog. https://blog.google/innovation-and-ai/infrastructure-and-cloud/google-cloud/cloud-next-2026-sundar-pichai/

110. Cybersecurity and Infrastructure Security Agency. (2025, January 14). CISA, JCDC, government and industry partners publish AI cybersecurity collaboration playbook. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/news/cisa-jcdc-government-and-industry-partners-publish-ai-cybersecurity-collaboration-playbook

111. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. https://doi.org/10.48550/arXiv.2509.22040

112. Chesney, R., & Citron, D. K. (2019). Deep fakes: A looming challenge for privacy, democracy, and national security. *California Law Review, 107*(6), 1753–1820. https://doi.org/10.15779/Z38RV0D

113. Schiff, K. J., Schiff, D. S., & Bueno, N. S. (2025). The liar's dividend: Can politicians claim misinformation to evade accountability?. *American Political Science Review*, *119*(1), 71-90.

114. Chan, Y.-T., et al. (2024). Assessing the article screening efficiency of artificial intelligence for systematic reviews. *Journal of Dentistry, 149*, 105259. https://doi.org/10.1016/j.jdent.2024.105259

115. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., Bennett, J. A., Jha, P., Mukhin, N., Li, J., López-Guajardo, E. A., Sadeghi, S., Bateni, F., & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. *Nature Chemical Engineering, 2*, 436–446. https://doi.org/10.1038/s44286-025-00249-z

116. Chan, A., Wei, K., Huang, S., Rajkumar, N., Perrier, E., Lazar, S., Hadfield, G. K., & Anderljung, M. (2025). Infrastructure for AI Agents. http://arxiv.org/abs/2501.10114

117. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. AI Agents That Matter. Transactions on Machine Learning Research.

118. Zhu, L., Lu, Q., Ding, M. et al. Designing meaningful human oversight in AI. AI Ethics 6, 286 (2026). https://doi.org/10.1007/s43681-026-01147-7

119. Ferrara, E. (2024). GenAI against humanity: Nefarious applications of generative artificial intelligence and large language models. *Journal of Computational Social Science, 7*, 549–569. https://doi.org/10.1007/s42001-024-00250-1

120. Djiré, A. E., Kaboré, A. K., Samhi, J., et al. (2026). *Learned or memorized? Quantifying memorization advantage in code LLMs*. In *Proceedings of the International Conference on Software Engineering*. https://arxiv.org/abs/2604.13997

121. Gowal, S., Bunel, R., Stimberg, F., Stutz, D., Ortiz-Jimenez, G., Kouridi, C., ... & Kohli, P. (2025). SynthID-Image: Image watermarking at internet scale. arXiv preprint arXiv:2510.09263.

122. United Nations Human Rights Council. Expert Mechanism on the Right to Development. (2024). *AI, cultural rights and the right to development* (A/HRC/EMRTD/11/CRP.2). United Nations. https://undocs.org/A/HRC/EMRTD/11/CRP.2

123. Brennan Center for Justice. (2025). *Gauging AI threat to free and fair elections*. https://www.brennancenter.org/our-work/analysis-opinion/gauging-ai-threat-free-and-fair-elections

124. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

125. Santos, F. P., Lelkes, Y., & Levin, S. A. (2021). Link recommendation algorithms and dynamics of polarization in online social networks. *Proceedings of the National Academy of Sciences*, *118*(50), e2102141118.

126. Cho, J., Ahmed, S., Hilbert, M., Liu, B., & Luu, J. (2020). Do search algorithms endanger democracy? An experimental investigation of algorithm effects on political polarization. *Journal of broadcasting & Electronic media*, *64*(2), 150-172.

127. Feezell, J. T., Wagner, J. K., & Conroy, M. (2021). Exploring the effects of algorithm-driven news sources on political behavior and polarization. *Computers in human behavior*, *116*, 106626.

128. Argyle, L. P. (2025). Political persuasion by artificial intelligence. *Science*, *390*(6777), 983-984.

129. Lin, H., Czarnek, G., Lewis, B., White, J. P., Berinsky, A. J., Costello, T., ... & Rand, D. G. (2025). Persuading voters using human–artificial intelligence dialogues. *Nature*, 1-8.

130. Myra Cheng et al. (2026). ,Sycophantic AI decreases prosocial intentions and promotes dependence.Science391,eaec8352(2026).DOI:10.1126/science.aec8352

131. Cheng, M., Lee, C., Khadpe, P., Yu, S., Han, D., & Jurafsky, D. (2026). Sycophantic AI decreases prosocial intentions and promotes dependence. Science, 391(6792), eaec8352. https://doi.org/10.1126/science.aec8352

132. Morrin, H., Nicholls, L., Levin, M., Yiend, J., Iyengar, U., DelGuidice, F., ... & Pollak, T. A. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. *The Lancet Psychiatry*, *13*(6), 522-530.

133. Balan, R., & Gumpel, T. P. (2025). ChatGPT Clinical Use in Mental Health Care: Scoping Review of Empirical Evidence. *JMIR Mental Health*, *12*, e81204.

134. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". *JMIR Mental Health*, *12*(1), e85799.

135. Osler, L. (2026). Hallucinating with AI: distributed delusions and "AI psychosis". Philosophy & Technology, 39(1), 30.

136. Askell, A., Bai, Y., Chen, A., Drain, D., Ganguli, D., Henighan, T., Jones, A., Joseph, N., Mann, B., DasSarma, N., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Kernion, J., Ndousse, K., Olsson, C., Amodei, D., Brown, T., Clark, J., ... Kaplan, J. (2021). A general language assistant as a laboratory for alignment (arXiv:2112.00861). arXiv. https://doi.org/10.48550/arXiv.2112.00861

137. Organisation for Economic Co-operation and Development. (2024). *Facts not fakes: Tackling disinformation, strengthening information integrity*. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

138. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. https://doi.org/10.1038/s41586-026-10506-7

139. Kalina Bontcheva (2024). Generative AI and Disinformation: Recent Advances, Challenges, and Opportunities. https://edmo.eu/wp-content/uploads/2023/12/Generative-AI-and-Disinformation_-White-Paper-v8.pdf

140. Laudrain, A. (2026, April 29). When AI governs (dis)information: Five lessons for democracy. Center for Security Studies (CSS), ETH Zurich. https://css.ethz.ch/en/center/CSS-news/2026/04/when-ai-governs-disinformation-five-lessons-for-democracy.html

141. Boine, C. (2023). Emotional attachment to AI companions and European law. *MIT Case Studies in Social and Ethical Responsibilities of Computing* (Winter 2023). https://doi.org/10.21428/2c646de5.db67ec7f

142. Department of Enterprise, Tourism and Employment. (2026, February 4). General scheme of the Regulation of Artificial Intelligence Bill 2026. Government of Ireland. https://www.gov.ie/en/department-of-enterprise-tourism-and-employment/publications/general-scheme-of-the-regulation-of-artificial-intelligence-bill-2026

143. United States Congress. Senate. (2025). *S. 3062—GUARD Act: Guidelines for User Age-verification and Responsible Dialogue Act of 2025* (119th Congress). Congress.gov. https://www.congress.gov/bill/119th-congress/senate-bill/3062/text

144. European Commission. (2026, April 29). *Blueprint for an age verification solution to help protect minors online*. Shaping Europe's Digital Future. https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online

145. Reuters. (2026, April 24). *From Australia to Europe, countries move to curb children's social media access*. https://www.reuters.com/legal/government/australia-europe-countries-move-curb-childrens-social-media-access-2026-04-24/

146. Morrin H, Nicholls L, Levin M et al. (2026). Artificial intelligence-associated delusions and large language models: risks, mechanisms of delusion co-creation, and safeguarding strategies. The Lancet Psychiatry, 2026; 13, 522-530

147. Examining the harm of AI chatbots, Hearing before the Subcomm. on Crime and Counterterrorism of the S. Comm. on the Judiciary, 119th Cong. (2025) (testimony of Megan Garcia). https://www.judiciary.senate.gov/download/2025-09-16-pm_-testimony_-garciapdf

148. Solove, D. J. (2025). Artificial intelligence and privacy. *Florida Law Review, 77*. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

149. Office of the United Nations High Commissioner for Human Rights. (2025). *The right to privacy in the digital age* (UN Doc. A/HRC/60/45). https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high

150. Bartlett, R., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 143*(1), 30–56. https://doi.org/10.1016/j.jfineco.2021.05.047

151. UNESCO, IRCAI, & University College London. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. UNESCO. https://unesdoc.unesco.org/ark:/48223/pf0000388971

152. Shah, S. S. (2024). Gender bias in artificial intelligence: Empowering women through digital literacy. *Premier Journal of Artificial Intelligence, 1*, 1000088. https://doi.org/10.70389/PJAI.1000088

153. Haider, S. A., Borna, S., Gomez-Cabello, C. A., et al. (2026). The algorithmic divide: A systematic review on AI-driven racial disparities in healthcare. *Journal of Racial and Ethnic Health Disparities, 13*, 188–217. https://doi.org/10.1007/s40615-024-02237-0

154. International Telecommunication Union. (2025). *Joint statement on AI and the rights of the child*. International Telecommunication Union. https://www.itu.int/hub/publication/d-str-cyb_joint-2025

155. Johnson, A. K., Winther, D. K., & Bhargava, A. (2026). *Artificial intelligence and child sexual exploitation and abuse: Emerging risks and implications for children's rights* (Issue brief). United Nations Children's Fund (UNICEF). https://www.unicef.org/media/178571/file/UNICEF%20AI%20CSEA%20Brief_FINAL3.pdf

156. Thiel, D. (2023). Identifying and Eliminating CSAM in Generative ML Training Data and Models. Stanford Digital Repository. Available at https://purl.stanford.edu/kh752sm9123

157. Internet Watch Foundation. (2026). Harm without limits: AI child sexual abuse material through the eyes of our Analysts. https://www.iwf.org.uk/media/hl1nvdti/iwf-ai-csam-report-2026.pdf

158. Goodacre, E.J. and Gibson, J.L. (2026) AI in the early years: Examining the implications of GenAI toys for young children. Cambridge: University of Cambridge (unpublished report, available via Apollo repository).

159. Kurian, N. (2025). Developmentally aligned AI: a framework for translating the science of child development into AI design. AI, Brain and Child, 1(1). https://doi.org/10.1007/s44436-025-00009-z

160. Livingstone, S., & Sylwander, K. R. (2025). Conceptualizing age-appropriate social media to support children's digital futures. British Journal of Developmental Psychology. https://doi.org/10.1111/bjdp.70006

161. Xiao, W., & Gonçalves, A. (2025). Intelligent toys, complex questions: A literature review of artificial intelligence in children's toys and devices. Big Data & Society, 12(4), 20539517251389860.

162. Grimmelikhuijsen, S. (2022). Explaining why the computer says no: Algorithmic transparency affects the perceived trustworthiness of automated decision-making. Public Administration Review, 82(4), 706–718. https://doi.org/10.1111/puar.13483

163. Richardson, R., Schultz, J. M., & Crawford, K. (2019). Dirty data, bad predictions: How civil rights violations impact police data, predictive policing systems, and justice. New York University Law Review Online, 94, 15–55. https://ssrn.com/abstract=3333423

164. Mantelero, A. (2022). Human rights impact assessment and AI. In Beyond data: Human rights, ethical and social impact assessment in AI (pp. 45-91). The Hague: TMC Asser Press.

165. Mantelero, A., & Esposito, M. S. (2021). An evidence-based methodology for human rights impact assessment (HRIA) in the development of AI data-intensive systems. Computer Law and Security Review, 41. https://doi.org/10.1016/j.clsr.2021.105561

166. United Nations Educational, Scientific and Cultural Organization. (2025). *How should children's rights be integrated into AI governance?* https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

167. Livingstone, S. & Pothong, K. (2025). Child Rights Impact Assessment: A Policy Tool for aRights-Respecting Digital Environment. Policy & Internet.

168. Denain, J.-S., & Barry, A. (2026, April 16). *Have AI capabilities accelerated?* Epoch AI. https://epoch.ai/blog/have-ai-capabilities-accelerated

169. Model Evaluation & Threat Research (METR). (2026, May 8). *Task-completion time horizons of frontier AI models*. https://metr.org/time-horizons/

170. Juniewicz, I. (2026, February 26). *Hyperscaler capex has quadrupled since GPT-4's release*. Epoch AI. https://epoch.ai/data-insights/hyperscaler-capex-trend

171. Epoch AI. (2026, May 28). *Data on AI companies.* https://epoch.ai/data/ai-companies?view=graph&tab=revenue&showTotal=true

172. Schölkopf, B., Locatello, F., Bauer, S., Ke, N. R., Kalchbrenner, N., Goyal, A., & Bengio, Y. (2021). Toward causal representation learning. Proceedings of the IEEE, 109(5), 612-634.

173. AI Alliance Network. (2025). AI Horizons: What will AI technologies look like in 10 years? A Research Project. November 2025, p. 83. Based on 21 foresight sessions and 32 in-depth interviews with over 270 AI researchers from 36 countries.

174. Gommers, J., Hernström, V., Josefsson, V., Sartor, H., Schmidt, D., Hjelmgren, A., ... & Lång, K. (2026). Interval cancer, sensitivity, and specificity comparing AI-supported mammography screening with standard double reading without AI in the MASAI study: a randomised, controlled, non-inferiority, single-blinded, population-based, screening-accuracy trial. The Lancet, 407(10527), 505-514.

175. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. *Nature Communications, 16*, 2564. https://www.nature.com/articles/s41467-025-57640-w

176. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. Scientific Reports, 15(1), 17458.

177. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K-12 education. npj Science of Learning, 10(1), 29.

178. Delgado-Licona, F., Alsaiari, A., Dickerson, H., Klem, P., Ghorai, A., Canty, R. B., ... & Abolhasani, M. (2025). Flow-driven data intensification to accelerate autonomous inorganic materials discovery. Nature Chemical Engineering, 2(7), 436-446.

179. Rotz, S., Duncan, E., Small, M., Botschner, J., Dara, R., Mosby, I., ... & Fraser, E. D. (2019). The politics of digital agricultural technologies: a preliminary review. Sociologia ruralis, 59(2), 203-229.

180. Afshar, M., Ryan Baumann, M., Resnik, F., Hintzke, J., Gravel Sullivan, A., Wills, G., ... & Gordon, J. (2025). A pragmatic randomized controlled trial of ambient artificial intelligence to improve health practitioner well-being. NEJM AI, 2(12), AIoa2500945.

181. Chen, M., Wu, Y., Ma, J., Jia, X., Gao, C., Zhao, F., & Qiao, Y. (2026). Independent and collaborative performance of large language models and healthcare professionals in diagnosis and triage. npj Digital Medicine.

182. Tao, X., Zhou, S., Ding, K., Li, S., Li, Y., Wu, B., ... & Han, S. (2026). An LLM chatbot to facilitate primary-to-specialist care transitions: a randomized controlled trial. Nature Medicine, 1-9.

183. Costa-Gomes, B., Tolmachev, P., Taysom, E., Sounderajah, V., Richardson, H., Schoenegger, P., ... & King, D. (2026). Public use of a generalist LLM chatbot for health queries. Nature Health, 1-8.

184. Scientific Computing World. (2025, August 5). *AI health assistant displays high diagnostic accuracy in tests*. https://www.scientific-computing.com/article/sber-healths-gigachat-powered-ai-health-assistant-displays-high-diagnostic-accuracy-tests

185. MASTEL, P. M., de Dieu Nyandwi, J., Rutunda, S., & Kabanda, K. (2025). Mbaza RBC: Deploying and evaluation of an LLM powered Chatbot for Community Health Workers in Rwanda. In *Workshop on Large Language Models and Generative AI for Health at AAAI 2025*.

186. Mateen, B. A., Williams, G., Korom, R., Mwaniki, P., Emmanuel-Fabula, M., & Agweyu, A. (2026). Learning Effects from A GenAI-based Clinical Decision Support System in Primary Healthcare. *medRxiv*, 2026-05.

187. OECD (2025), Results from TALIS 2024: The State of Teaching, TALIS, OECD Publishing, Paris, https://doi.org/10.1787/90df6235-en.

188. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en .

189. Létourneau, A., Deslandes Martineau, M., Charland, P., Karran, J. A., Boasen, J., & Léger, P. M. (2025). A systematic review of AI-driven intelligent tutoring systems (ITS) in K–12 education. *npj Science of Learning, 10*(1), 29. https://www.nature.com/articles/s41539-025-00320-7

190. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

191. OECD (2023), PISA 2022 Results (Volume II): Learning During – and From – Disruption, PISA, OECD Publishing, Paris, https://doi.org/10.1787/a97db61c-en.

192. Vodafone Foundation.(2025) AI in European Schools: A European Report --- comparing seven countries

193. World Food Programme. (2024). *HungerMap LIVE: Global hunger monitoring*. https://hungermap.wfp.org/

194. Yakov and Partners. (2024). *Artificial intelligence in Russia's agricultural sector: Hype or real money?* https://yakovpartners.com/publications/ai-in-agriculture/

195. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2025). Generative AI without guardrails can harm learning: Evidence from high school mathematics. *Proceedings of the National Academy of Sciences, 122*(26), e2422633122. https://doi.org/10.1073/pnas.2422633122

196. Kestin, G., Miller, K., Klales, A., Milbourne, T., & Ponti, G. (2025). AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports, 15*, 17458. https://doi.org/10.1038/s41598-025-97652-6

197. Shneiderman, B. (2022). Human-Centered AI. Oxford University Press.

198. Sparling, T. M., Offner, C., Deeney, M., Denton, P., Bash, K., Juel, R., ... & Kadiyala, S. (2024). Intersections of climate change with food systems, nutrition, and health: an overview and evidence map. Advances in Nutrition, 15(9), 100274.

199. Reichstein, M., et al. (2025). Early warning of complex climate risk with integrated artificial intelligence. Nature Communications, 16, 2564. https://doi.org/10.1038/s41467-025-57640-w

200. Becker-Reshef, I., Justice, C., Barker, B., Humber, M., Rembold, F., Bonifacio, R., Zappacosta, M., Budde, M., Magadzire, T., Shitote, C., Pound, J., Constantino, A., Nakalembe, C., Mwangi, K., Sobue, S., Newby, T., Whitcraft, A., Jarvis, I., & Verdin, J. (2020). Strengthening agricultural decisions in countries at risk of food insecurity: The GEOGLAM Crop Monitor for Early Warning. *Remote Sensing of Environment, 237*, 111553. https://doi.org/10.1016/j.rse.2019.111553

201. Food and Agriculture Organization of the United Nations. (2025). *Evidence in action: How anticipatory cash transfers reduce humanitarian needs and strengthen resilience in Somalia*. https://www.fao.org/agrifood-economics/publications/detail/en/c/1756210/

202. World Food Programme. (2025). *WFP's evidence base on anticipatory action 2015–2024*. https://www.wfp.org/publications/wfps-evidence-base-anticipatory-action-2015-2024

203. Nakalembe, C., Kerner, H. R., Zvonkov, I., Humber, M., Galvez, A. S., Venturini, S., & Becker Reshef, I. (2025). A framework for EO based National Agricultural Monitoring for the African context. npj Sustainable Agriculture, 3, 45. https://www.nature.com/articles/s44264-025-00083-z

204. Nowak, A. C., et al. (2024). Opportunities to strengthen Africa's efforts to track national level climate adaptation. Nature Climate Change, 14, 876–882.https://www.nature.com/articles/s41558-024-02054-7

205. Karger, E., Kuusela, O., Abaluck, J., Bryan, K. A., Halperin, B., Jones, T. R., et al. (2026). *Forecasting the economic effects of AI* (NBER Working Paper No. w35046). National Bureau of Economic Research. https://doi.org/10.3386/w35046

206. Goldman Sachs (2023). Generative AI Could Raise Global GDP by 7 Percent. https://www.goldmansachs.com/insights/articles/generative-ai-could-raise-global-gdp-by-7-percent

207. Anantrasirichai, N., & Bull, D. (2022). Artificial intelligence in the creative industries: a review. Artificial intelligence review, 55(1), 589-656.

208. United Nations News. (2026, February 18). *Artists face steep income decline due to AI, UNESCO finds*. United Nations. https://news.un.org/en/story/2026/02/1166989

209. Brynjolfsson, E., Rock, D., & Syverson, C. (2021). The productivity J-curve: How intangibles complement general purpose technologies. American Economic Journal: Macroeconomics, 13(1), 333-372.

210. Gmyrek, P., Winkler, H., & Garganta, S. (2024). Buffer or Bottleneck? Employment Exposure to Generative AI and the Digital Divide in Latin America (ILO Working Paper 121 / World Bank Policy Research Working Paper 10863). International Labour Organization & World Bank.

211. Autor, D., Chin, C., Salomons, A., & Seegmiller, B. (2024). New frontiers: The origins and content of new work, 1940–2018. Quarterly Journal of Economics, 139(3), 1399–1465.

212. The Conversation. (2026, March 18). *Tech companies are blaming massive layoffs on AI. What's really going on?* https://theconversation.com/tech-companies-are-blaming-massive-layoffs-on-ai-whats-really-going-on-278314

213. Society for Human Resource Management. (2026, May). *The AI layoffs narrative: Real transformation, or scapegoat?* https://www.shrm.org/topics-tools/news/technology/ai-layoffs-transformation-scapegoat

214. OpenAI. (2026, February 27). Company communication accompanying a US$110 billion private funding round [Company communication]

215. Rodrik, D., & Sabel, C. (2020). Building a Good Jobs Economy (HKS Faculty Research Working Paper RWP20-001). Harvard Kennedy School.

216. Acemoglu, D. (2025). The simple macroeconomics of AI. Economic Policy, 40(121), 13–58. Acemoglu's headline figure is a TFP gain of no more than 0.71% over ten years.

217. Korinek, A., & Suh, D. (2024). Scenarios for the Transition to AGI (NBER Working Paper No. 32255). In their full-automation scenario, output rises sharply while wages collapse once automation crosses a critical threshold.

218. Karger, E., Kuusela, O., Abaluck, J., Bryan, K., Halperin, B., Jones, T., et al. (2026). Forecasting the Economic Effects of AI. Federal Reserve Bank of Chicago and Forecasting Research Institute, March 2026.

219. Jones, C. I., & Tonetti, C. (2026). Past Automation and Future AI: How Weak Links Tame the Growth Explosion. Working paper presented at the Bendheim Center for Finance, Princeton University, March 2026.

220. Trammell, P., & Korinek, A. (2025). Economic Growth under Transformative AI (NBER Working Paper No. 31815, revised September 2025).

221. OECD (2024). Artificial Intelligence, Data and Competition (OECD Artificial Intelligence Papers No. 18). OECD Publishing, Paris. https://doi.org/10.1787/e7e88884-en

222. Acemoglu, D., & Restrepo, P. (2026). Automation and rent dissipation: Implications for wages, inequality, and productivity. Quarterly Journal of Economics, 141(2), 1521–1579.

223. Liu, Y., Zhao, Y., Lyu, Y., Zhang, T., Wang, H., & Lo, D. (2025). "Your AI, my shell": Demystifying prompt injection attacks on agentic AI coding editors. arXiv. https://doi.org/10.48550/arXiv.2509.22040

224. Regilme, S. S. F. (2024). Artificial Intelligence Colonialism: Environmental Damage, Labor Exploitation, and Human Rights Crises in the Global South. SAIS Review of International Affairs, 44(2), 75–92. https://muse.jhu.edu/pub/1/article/950958

225. Barnett-Itzhaki, Z. (2026). The water footprint of artificial intelligence: Emerging solutions and governance imperatives. Water Research, 299, 125866. https://doi.org/10.1016/j.watres.2026.125866

226. International Energy Agency. (2025). Global Critical Minerals Outlook 2025 – Analysis (p. 312). https://iea.blob.core.windows.net/assets/ef5e9b70-3374-4caa-ba9d-19c72253bfc4/GlobalCriticalMineralsOutlook2025.pdf

227. Zhu, L., & Lu, Q. (2026). Verifiability-First AI Engineering in the Era of AIware: A Conceptual Framework, Design Principles, and Architectural Patterns for Scalable Verification. Design Principles, and Architectural Patterns for Scalable Verification (January 07, 2026).

228. UN Scientific Advisory Board. (2026, March 19). AI deception: Brief of the Scientific Advisory Board. United Nations. https://www.un.org/scientific-advisory-board/en/ai-deception

229. Bengio, Y., Clare, S., Prunkl, C., Rismani, S., Andriushchenko, M., Bucknall, B., ... & Zhu, L. (2025). International AI Safety Report 2025: First Key Update: Capabilities and Risk Implications. *arXiv preprint arXiv:2510.13653*.

230. United Nations Secretary-General. (2024). *Human rights due diligence guidance on digital technology use*. https://www.ohchr.org/sites/default/files/2024-08/digital-technology-use-guidance-sg-1-en.pdf

231. AI Equality Toolbox. (2025). *Human rights impact assessment methodology*. https://aiequalitytoolbox.com/tools/hria-workbook/

232. Baldé, C. P., Kuehr, R., Yamamoto, T., McDonald, R., D'Angelo, E., Althaf, S., Bel, G., Deubzer, O., Fernandez-Cubillo, E., Forti, V., Gray, V., Herat, S., Honda, S., Iattoni, G., Khetriwal, D. S., Luda di Cortemiglia, V., Lobuntsova, Y., Nnorom, I., Pralat, N., & Wagner, M. (2024). *The global e-waste monitor 2024: Electronic waste rising five times faster than documented e-waste recycling*. United Nations Institute for Training and Research & International Telecommunication Union. https://ewastemonitor.info/the-global-e-waste-monitor-2024/

233. International Panel on the Information Environment. (2024). *Expert survey on the global information environment 2024: Searching for solutions (SFP2024-1)*. https://www.ipie.info/research/sfp2024-1

234. Aïmeur, E., Amri, S., & Brassard, G. (2023). Fake news, disinformation and misinformation in social media: A review. *Social Network Analysis and Mining, 13*(1), 30. https://doi.org/10.1007/s13278-023-01028-5

235. James, K., Perveen, S., & Jacobson, B. (2025). *Drained by data: The cumulative impact of data centers on regional water stress*. Ceres. https://www.ceres.org/resources/reports/drained-by-data-the-cumulative-impact-of-data-centers-on-regional-water-stres

236. Office of the United Nations High Commissioner for Human Rights. (2024). *Taxonomy of generative AI-related human rights harms*. https://www.ohchr.org/sites/default/files/documents/issues/expression/statements/2025-10-24-joint-declaration-artificial-intelligence.pdf

237. Patel, A. (2025, May 19). *Freedom of expression, artificial intelligence and elections*. United Nations Educational, Scientific and Cultural Organization (UNESCO) & United Nations Development Programme (UNDP). https://www.undp.org/publications/freedom-expression-artificial-intelligence-and-elections

238. Scharfbillig, M., Lewandowsky, S., Altay, S., Van Alstyne, M., Kozyreva, A., Hertwig, R., Lorenz-Spreen, P., DiResta, R., Valenzuela, S., Egidy, S., Quattrocioc-chi, W., & Orben, A. (2026). *Fractured reality: How democracy can win the global struggle over the information space* (JRC144603). Publications Office of the European Union. https://doi.org/10.2760/9358883

239. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

240. Solove, D. J. (2025). *Artificial intelligence and privacy*. *Florida Law Review, 77*. https://scholarship.law.ufl.edu/flr/vol77/iss1/1

241. Office of the United Nations High Commissioner for Human Rights (2025). The right to privacy in the digital age. URL https://www.ohchr.org/en/documents/thematic-reports/ahrc6045-right-privacy-digital-age-reportoffice-united-nations-high. UN Doc. A/HRC/60/45.

242. Council of Europe. Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

243. European Union. (2024, June 13). *Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. *Official Journal of the European Union, L series*. https://eur-lex.europa.eu/eli/reg/2024/1689/oj

244. Expert Mechanism on the Right to Development. (2024). *Artificial intelligence, cultural rights and the right to development* (A/HRC/EMRTD/13/CRP.1). United Nations Human Rights Council. https://www.ohchr.org/sites/default/files/documents/issues/development/emd/session13/a-hrc-emrtd-13-crp-1.pdf

245. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

246. Waight, H., Yang, E., Yuan, Y., et al. (2026). State media control influences large language models. *Nature*. https://doi.org/10.1038/s41586-026-10506-7

247. Li, P., Yang, J., Islam, M. A., & Ren, S. (2025). Making AI less "thirsty": Uncovering and addressing the secret water footprint of AI models. *Communications of the ACM, 68*(7), 54–61. https://doi.org/10.1145/3724499

248. Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S., ... & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. arXiv preprint arXiv:2508.15734.

249. Arsenault, A. C., & Kreps, S. (2026). Whose voice counts? The role of large language models in public commenting. *Big Data & Society, 13*(1). https://doi.org/10.1177/20539517261419341

250. Alslaity, A., Chan, G., & Orji, R. (2023). A panoramic view of personalization based on individual differences in persuasive and behavior change interventions. *Frontiers in Artificial Intelligence, 6*, 1125191. https://doi.org/10.3389/frai.2023.1125191

251. Costello, T. H., Pennycook, G., & Rand, D. G. (2024). Durably reducing conspiracy beliefs through dialogues with AI. *Science, 385*, eadq1814. https://doi.org/10.1126/science.adq1814

252. Boissin, E., Costello, T. H., Spinoza-Martín, D., Rand, D. G., & Pennycook, G. (2025). Dialogues with large language models reduce conspiracy beliefs even when the AI is perceived as human. *PNAS Nexus, 4*(11), pgaf325. https://doi.org/10.1093/pnasnexus/pgaf325

253. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

254. Hackenburg, K., Tappin, B. M., Hewitt, L., Saunders, E., Black, S., Lin, H., Fist, C., Margetts, H., Rand, D. G., & Summerfield, C. (2025). The levers of political persuasion with conversational AI. Science, 390(6777), eaea3884. https://doi.org/10.1126/science.aea3884

255. Germano, F., Gómez, V., & Sobbrio, F. (2025). *Ranking for engagement: How social media algorithms fuel misinformation and polarization*. Barcelona School of Economics, Working Paper No. 1501. https://bw.bse.eu/wp-content/uploads/2025/07/1501.pdf

256. Council of Europe CDMSI. (2025). Guidance Note on Generative AI and Freedom of Expression. CDMSI(2025) https://rm.coe.int/cdmsi-2025-15rev-guidance-note-on-the-implications-of-generative-artif/488029df80

257. Buyl, M., Rogiers, A., Noels, S., Bied, G., Dominguez-Catena, I., Heiter, E., ... & De Bie, T. (2026). Large language models reflect the ideology of their creators. *npj Artificial Intelligence*, *2*(1), 7.

258. Wang, P., Zhang, L.-Y., Tzachor, A., & Chen, W.-Q. (2024). E-waste challenges of generative artificial intelligence. Nature Computational Science, 4, 818–823. https://doi.org/10.1038/s43588-024-00700-3

259. Schroeder, D. T., Cha, M., Baronchelli, A., Bostrom, N., Christakis, N. A., Garcia, D., ... & Kunst, J. R. (2026). How malicious AI swarms can threaten democracy. Science, 391(6783), 354-357.

260. Council of Europe. (2024). *Council of Europe framework convention on artificial intelligence and human rights, democracy and the rule of law* (Council of Europe Treaty Series No. 225). https://www.coe.int/en/web/artificial-intelligence/the-framework-convention-on-artificial-intelligence

261. Observatory on Information and Democracy. (2024). *Information ecosystems and troubled democracy: A global synthesis of the state of knowledge on news, media, AI, and data governance*. https://observatory.informationdemocracy.org/report/information-ecosystem-and-troubled-democracy/

262. Council of Europe, Steering Committee on Media and Information Society. (2025). *Guidance note on the implications of generative artificial intelligence for freedom of expression* (CDMSI(2025)15rev). https://rm.coe.int/cdmsi-2025-15rev

263. OECD. (2024). Facts not fakes: Tackling disinformation, strengthening information integrity. OECD Publishing. https://doi.org/10.1787/d909ff7a-en

264. United Nations General Assembly. (2017). *Promotion and protection of human rights: Human rights questions, including alternative approaches for improving the effective enjoyment of human rights and fundamental freedoms: Report of the Third Committee, 72nd session* (A/72/...). United Nations Digital Library. http://digitallibrary.un.org/record/1326669

265. Penney, J. W. (2025). *Chilling effects: Repression, conformity, and power in the digital age*. Cambridge University Press. https://doi.org/10.1017/9781108918022

266. UN Women. (2022). Tipping point: The chilling escalation of online violence against women in the public sphere. UN Women. https://www.unwomen.org/sites/default/files/2025-12/tipping-point-the-chilling-escalation-of-violence-against-women-in-the-public-sphere-in-the-age-of-ai-en.pdf

267. United Nations Educational, Scientific and Cultural Organization. (2024). *Challenging systematic prejudices: An investigation into bias against women and girls in large language models*. https://unesdoc.unesco.org/ark:/48223/pf0000388971

268. Chowdhury, R., & Lakshmi, D. (2023). “Your opinion doesn’t matter, anyway”: Exposing technology-facilitated gender-based violence in an era of generative AI (2nd ed.). UNESCO.

269. United Nations Educational, Scientific and Cultural Organization. (2024, March 7). *Generative AI: UNESCO study reveals alarming evidence of regressive gender stereotypes*. https://www.unesco.org/en/articles/generative-ai-unesco-study-reveals-alarming-evidence-regressive-gender-stereotypes

270. Fung, P. (2019, June 30). *This is why AI has a gender problem*. World Economic Forum. https://www.weforum.org/stories/2019/06/this-is-why-ai-has-a-gender-problem/

271. United Nations Conference on Trade and Development. (2025). *Technology and innovation report 2025: The AI divide*. https://unctad.org/publication/technology-and-innovation-report-2025

272. Ahmed, N., & Wahed, M. (2020). The De-democratization of AI: Deep learning and the compute divide in artificial intelligence research. *arXiv preprint arXiv:2010.15581*.

273. Coeckelbergh, M. (2026) Technofascism: AI, Big Tech, and the New Authoritarianism. AI & Society https://doi.org/10.1007/s00146-026-02862-9

274. Varoufakis, Y. (2024). *Technofeudalism: What killed capitalism*. Melville House.

275. Kalluri, P. R., Agnew, W., Cheng, M., Owens, K., Soldaini, L., & Birhane, A. (2025). Computer-vision research powers surveillance technology. Nature, 643(8070), 73-79. https://www.nature.com/articles/s41586-025-08972-6

276. Fola-Rose, A., Solomon, E., Bryant, K., & Woubie, A. (2024, August). A systematic review of facial recognition methods: Advancements, applications, and ethical dilemmas. In *Proceedings of the 2024 IEEE International Conference on Information Reuse and Integration for Data Science (IRI 2024)* (pp. 314–319). IEEE. https://doi.org/10.1109/IRI62200.2024.00070

277. Fussey, P., & Murray, D. (2025). Facial Recognition Surveillance: Policing and Human Rights in the Age of Artificial Intelligence. Oxford University Press.

278. Office of the United Nations High Commissioner for Human Rights. (2024). *Mapping report: Human rights and new and emerging digital technologies* (A/HRC/56/45). https://www.ohchr.org/en/documents/reports/mapping-report-human-rights-and-new-and-emerging-digital-technologies

279. Secretary-General. (2024). Human rights in the administration of justice (A/79/296). United Nations. https://digitallibrary.un.org

280. American Civil Liberties Union, More than a Dozen Wrongful Arrests Due to Police Reliance on Facial Recognition Technology (2025), https://www.aclu.org/news/privacy-technology/more-than-a-dozen-wrongful-arrests-due-to-police-reliance-on-facial-recognition-technology: Documentation of at least 14 publicly known wrongful arrests in the United States attributed to police use of facial recognition; in nearly all cases the persons wrongfully arrested were Black.

281. Saxena, D., & Guha, S. (2024). Algorithmic harms in child welfare: Uncertainties in practice, organization, and street-level decision-making. *ACM Journal on Responsible Computing, 1*(1), Article 2, 1–32. https://doi.org/10.1145/3616473

282. German Marshall Fund. (2024). Spitting Images: Tracking Deepfakes and Generative AI in Elections.

283. International Institute for Democracy and Electoral Assistance. (2024). *The 2024 global elections super-cycle*. https://www.idea.int/initiatives/the-2024-global-elections-supercycle

284. Recorded Future. (2024). 2024 Deepfakes and Election Disinformation Report: Key Findings and Mitigation Strategies.

285. Associated Press. (2025, June 13). New Hampshire jury acquits consultant behind AI robocalls mimicking Biden on all charges.

286. Federal Communications Commission. (2024, September). $6 million fine against Steven Kramer for AI-generated robocalls.

287. Global Witness. (2024). What Happened on TikTok Around the Romanian Elections?

288. IFES. (2024). The Romanian 2024 Election Annulment: Addressing Emerging Threats to Electoral Integrity.

289. Associated Press. (2024). Election disinformation takes a big leap with AI being used to deceive worldwide.

290. United Nations. (1966). International Covenant on Civil and Political Rights. Articles 17, 18, 19 and 25.

291. Council of Europe. (1950). Convention for the Protection of Human Rights and Fundamental Freedoms. Articles 8, 9 and 10; Protocol No. 1, Article 3.

292. EQUATE Language AI Readiness Index (https://equate.vercel.app/en)

293. Han, W., Zhang, Y., Chen, Z., Liu, B., Lin, H., Zhang, B., ... & Zheng, Y. (2025). MuBench: Assessment of Multilingual Capabilities of Large Language Models Across 61 Languages. arXiv preprint arXiv:2506.19468.

294. Lissak, S., Calderon, N., Shenkman, G., Ophir, Y., Fruchter, E., Klomek, A. B., & Reichart, R. (2024, June). The colorful future of llms: Evaluating and improving llms as emotional supporters for queer youth. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 2040-2079).

295. Gamboa, L. C. L., Feng, Y., & Lee, M. (2025, November). Social Bias in Multilingual Language Models: A Survey. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 27845-27868).

296. Choudhury, M., Sitaram, S., Vashistha, A., et al. (2026). *AI for the Global South: 12 critical research questions for the next decade*. AI for the Global South (AI4GS), Mohamed bin Zayed University of Artificial Intelligence. https://ai4gs.github.io/

297. Bartl, M., Mandal, A., Leavy, S., & Little, S. (2025). Gender bias in natural language processing and computer vision: A comparative survey. ACM Computing Surveys, 57(6), 1-36. https://dl.acm.org/doi/pdf/10.1145/3700438

298. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/sites/default/files/research/report/talk-trust-and-trade-offs_2025_web.pdf

299. U.S. PIRG Education Fund. (2025). AI comes to playtime: Artificial companions, real risks. U.S. PIRG Education Fund. https://pirg.org/edfund/wp-content/uploads/2025/12/AI-Comes-to-Playtime-Artifical-companions-real-risks.pdf

300. Staksrud, E., Mascheroni, G., Milosevic, T., Ní Bhroin, N., Ólafsson, K., Şengül-İnal, G., & Stoilova, M. (2026). European children's use and understanding of generative AI. EU Kids Online V.

301. Committee on the Rights of the Child. (2021). *General comment No. 25 (2021) on children's rights in relation to the digital environment* (CRC/C/GC/25). United Nations. https://digitallibrary.un.org/record/3906061

302. UNICEF (2025). Guidance on AI and Children: Updated guidance for governments and businesses to create AI policies and systems that uphold children's rights. https://www.unicef.org/innocenti/media/11991/file/UNICEF-Innocenti-Guidance-on-AI-and-Children-3-2025.pdf

303. UNESCO (2025). How should children's rights be integrated into AI governance? https://www.unesco.org/en/articles/how-should-childrens-rights-be-integrated-ai-governance

304. Grossman, S., Pfefferkorn, R., & Liu, S. (2025). AI-Generated Child Sexual Abuse Material: Insights from Educators, Platforms, Law Enforcement, Legislators, and Victims. Version 1. Stanford Digital Repository. Available at https://purl.stanford.edu/mn692xc5736/version/1. https://doi.org/10.25740/mn692xc5736.

305. Livingstone, S., Atabey, A., Stoilova, M., & Sylwander, K. R. (2025). How does, and how could, generative AI respect and enable children's rights? In N. Ni Loideain (Ed.), *AI and power: Regulation and rights*. University of London Press.

306. European Parliamentary Research Service. (2024). Children and deepfakes. European Parliament. https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2025)775855

307. United Nations Children's Fund (UNICEF). (2026). *Artificial intelligence and child sexual abuse and exploitation* [Issue brief]. https://www.unicef.org/reports/artificial-intelligence-and-child-sexual-abuse-and-exploitation

308. Ozcan, B., Sperati, V., Giocondo, F., Schembri, M., & Baldassarre, G. (2022, June). Interactive soft toys to support social engagement through sensory-motor plays in early intervention of kids with special needs. In Proceedings of the 21st Annual ACM Interaction Design and Children Conference (pp. 625-628).

309. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://www.cam.ac.uk/stories/ai-toys-study-play

310. British Standards Institution. (2026, May). *Half of children have AI toys despite safety concerns*. https://www.bsigroup.com/en-GB/insights-and-media/media-centre/press-releases/2026/may/half-of-children-have-ai-toys-as-parents-allow-widespread-use-despite-safety-concerns-and-gaps-in-guidance-parents/

311. Goodacre, E., & Gibson, J. (2026). AI in the Early Years: Examining the implications of GenAI toys for young children. https://doi.org/10.17863/CAM.126270

312. Chou, C. Y., Chan, T. W., Chen, Z. H., Liao, C. Y., Shih, J. L., Wu, Y. T., & Hung, H. C. (2025). Defining AI companions: a research agenda--from artificial companions for learning to general artificial companions for Global Harwell. *Research & Practice in Technology Enhanced Learning*, *20*.

313. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic artificial intelligence companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

314. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications. Leverhulme Centre for the Future of Intelligence

315. De Freitas, J., Oguz-Uguralp, Z., & Kaan-Uguralp, A. (2025). Emotional manipulation by AI companions. arXiv preprint arXiv:2508.19258.

316. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

317. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

318. Zhang, R., Li, H., Meng, H., Zhan, J., Gan, H., & Lee, Y. C. (2025, April). The dark side of ai companionship: A taxonomy of harmful algorithmic behaviors in human-ai relationships. In Proceedings of the 2025 CHI conference on human factors in computing systems (pp. 1-17).

319. De Freitas, J., & Cohen, I. G. (2024). The health risks of generative AI-based wellness apps. Nature medicine, 30(5), 1269-1275.

320. Radesky, J., Bragg, M. A., & Hiniker, A. (2026). Risks and Consequences of Children's Use of Social AI—A Framework. JAMA Pediatrics. https://doi.org/10.1001/jamapediatrics.2026.1349

321. Muldoon, J., & Parke, J. J. (2025). Cruel companionship: How AI companions exploit loneliness and commodify intimacy. new media & society, 14614448251395192.

322. Jacobs, K. A. (2024). Digital loneliness—changes of social recognition through AI companions. Frontiers in Digital Health, 6, 1281037.

323. Ho, J. Q., Hu, M., Chen, T. X., & Hartanto, A. (2025). Potential and pitfalls of romantic Artificial Intelligence (AI) companions: A systematic review. Computers in Human Behavior Reports, 19, 100715.

324. Hollanek, T., & Sobey, A. (2025). AI companions for health and mental wellbeing: opportunities, risks and policy implications.

325. Zhang, Y., Zhao, D., Hancock, J. T., Kraut, R., & Yang, D. (2025). The rise of AI companions: how human-chatbot relationships influence well-being. arXiv preprint arXiv:2506.12605.

326. Dewitte, P. (2024). Better alone than in bad company: Addressing the risks of companion chatbots through data protection by design. Computer Law & Security Review, 54, 106019.

327. Robb, M. B., & Mann, S. (2025). *Talk, trust, and trade-offs: How and why teens use AI companions*. Common Sense Media. https://www.commonsensemedia.org/research/talk-trust-and-trade-offs-how-and-why-teens-use-ai-companions

328. Rousmaniere, T., Zhang, Y., Li, X., & Shah, S. (2025). Large language models as mental health resources: Patterns of use in the United States. *Practice Innovations*. https://doi.org/10.1037/pri0000292

329. Callahan, C., Tanner, L., Coe, C., Davis, M., Glover, J., Bernstein, E., ... & Kunkle, S. (2026). Real-World Use of a Mental Health AI Companion: Multiple Methods Study. JMIR Formative Research, 10, e86904.

330. Associated Press. (2026, June 1). *Chatbot AI lawsuit alleges links to teen suicide*. https://apnews.com/article/chatbot-ai-lawsuit-suicide-teen-artificial-intelligence-9d48adc572100822fdbc3c90d1456bd0

331. Hudon, A., & Stip, E. (2025). Delusional experiences emerging from AI chatbot interactions or "AI Psychosis". JMIR Mental Health, 12(1), e85799.

332. Green, H. H. (2026, March 14). *New study raises concerns about AI chatbots fueling delusional thinking*. The Guardian. https://www.theguardian.com/technology/2026/mar/14/ai-chatbots-psychosis

333. Ministère du Travail, de la Santé, des Solidarités et des Familles. (2025, March 24). *La santé mentale, grande cause nationale 2025*. Gouvernement de la France. https://solidarites.gouv.fr/la-sante-mentale-grande-cause-nationale-2025

334. American Psychiatric Association. (n.d.). *Applications of artificial intelligence in mental health care*. https://www.psychiatry.org/psychiatrists/practice/artificial-intelligence/applications

335. U.S. Food and Drug Administration. (2025). *Executive summary for the Digital Health Advisory Committee meeting: Generative artificial intelligence–enabled digital mental health medical devices*. https://www.fda.gov/media/189391/download

336. Hollis, A., & McKeown, G. (2024, September). Empathic AI for autism: Potential and pitfalls of empathic social chatbots in addressing loneliness. In 24th ACM International Conference on Intelligent Virtual Agents: CONNECT, A Workshop on Connecting Interdisciplinary Research on Connections With and Through Technology: IVA 2024.

337. Sharma, D., Meshkat, S., Perivolaris, A., Kamaleddin, M. A., Teferra, B. G., Rueda, A., ... & Bhat, V. (2026). Reimagining psychiatric care with agentic AI: promise, challenges, and a roadmap forward. npj Digital Medicine.

338. Straw, I., & Callison-Burch, C. (2020). Artificial Intelligence in mental health and the biases of language based models. PloS one, 15(12), e0240376.

339. Garg, M. (2024). Towards mental health analysis in social media for low-resourced languages. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(3), 1-22.

340. Wang, W., Tu, Z., Chen, C., Yuan, Y., Huang, J.-T., Jiao, W., & Lyu, M. R. (2024). All languages matter: On the multilingual safety of LLMs. *Findings of the Association for Computational Linguistics: ACL 2024*, 5865–5877. https://doi.org/10.18653/v1/2024.findings-acl.349

341. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). *Viability of machine translation for healthcare in low-resourced languages*. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

342. Fu, Y. V., Ramachandran, G. K., Park, N., Lybarger, K., Xia, F., Uzuner, Ö., & Yetisgen, M. (2025). BioMistral-NLU: Towards more generalizable medical language understanding through instruction tuning. *AMIA Joint Summits on Translational Science Proceedings, 2025*, 149–158. https://pubmed.ncbi.nlm.nih.gov/40502228/

343. Labrak, Y., Bazoge, A., Morin, E., Gourraud, P.-A., Rouvier, M., & Dufour, R. (2024). BioMistral: A collection of open-source pretrained large language models for medical domains. *Findings of the Association for Computational Linguistics: ACL 2024*, 5848–5864. https://aclanthology.org/2024.findings-acl.348/

344. Qiu, P., Wu, C., Zhang, X., Lin, W., Wang, H., Zhang, Y., Wang, Y., & Xie, W. (2024). Towards building multilingual language models for medicine. *Nature Communications, 15*, 8384. https://doi.org/10.1038/s41467-024-52417-z

345. Nwabufo, J., Ogueji, K., Adelani, D. I., Alabi, J., et al. (2025). Healthcare NLP for African Languages: Current State and Challenges. Proceedings of the AfricaNLP Workshop (AfricaNLP 2025). https://aclanthology.org/2025.africanlp-1.32/

346. Okafor, U. (2025). Multilingual NLP for African Healthcare: Bias, Translation, and Explainability Challenges. Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025), 221–229. https://aclanthology.org/2025.africanlp-1.32/

347. Skianis, K., Doğruöz, A. S., & Pavlopoulos, J. (2024). Leveraging LLMs for translating and classifying mental health data. In J. Sälevä & A. Owodunni (Eds.), Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024) (pp. 236–241). https://doi.org/10.18653/v1/2024.mrl-1.20

348. Cronin, A., Kelly, A., Wrona, M., O'Donnell, P., Hassan, A., Myles, T., Fallon, T., & MacFarlane, A. (2025). The patient-safety implications of AI-based communication with migrants in general practice: a scoping review. BJGP Open, 9(4), BJGPO.2025.0107. https://bjgpopen.org/content/9/4/BJGPO.2025.0107

349. House of Lords Public Services Committee. (2025). *Lost in translation? Interpreting services in the courts* (2nd Report of Session 2024–26, HL Paper 87). https://committees.parliament.uk/publications/44602/documents/221328/default/

350. Nigatu, H. H., Mehandru, N., Abadi, N. H., Gebremeskel, B., Alaa, A., & Choudhury, M. (2025). Viability of machine translation for healthcare in low-resourced languages. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, 10584–10598. https://aclanthology.org/2025.emnlp-main.535/

351. MIT AI Risk Initiative. (2025). *AI risk mitigation database and draft taxonomy*. https://airisk.mit.edu/ai-riskmitigations

352. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244*.

353. Staufer, L., Feng, K., Wei, K., et al. (2026). *The 2025 AI agent index: Documenting technical and safety features of deployed agentic AI systems*. https://doi.org/10.48550/arXiv.2602.17753

354. Weidinger, L., Raji, I. D., Wallach, H., Mitchell, M., Wang, A., Salaudeen, O., ... & Isaac, W. (2025). Toward an evaluation science for generative ai systems. *arXiv preprint arXiv:2503.05336*.

355. Rabanser, S., Kapoor, S., Kirgis, P., et al. (2026). *Towards a science of AI agent reliability*. https://doi.org/10.48550/arXiv.2602.16666

356. Bastani, H., Bastani, O., Sungu, A., Ge, H., Kabakcı, Ö., & Mariman, R. (2024). Generative AI can harm learning. *The Wharton School Research Paper*.

357. Shen, J. H., & Tamkin, A. (2026). How AI impacts skill formation. *arXiv preprint arXiv:2601.20245*.

358. Budzyń, K., Romańczyk, M., Kitala, D., Kołodziej, P., Bugajski, M., Adami, H. O., ... & Mori, Y. (2025). Endoscopist deskilling risk after exposure to artificial intelligence in colonoscopy: a multicentre, observational study. *The Lancet Gastroenterology & Hepatology*, *10*(10), 896-903.

359. Epoch AI. (2026). *Data on AI models*. https://epoch.ai/data/ai-models

360. Feng, K. J., McDonald, D. W., & Zhang, A. X. (2025). Levels of autonomy for ai agents. *arXiv preprint arXiv:2506.12469*.

361. Fink, M. (2025). *Operationalizing meaningful human oversight under Article 14 of the EU AI Act*. In *AI Act commentary: A thematic analysis* (forthcoming). Hart-Bloomsbury.

362. Shapira, N., Wendler, C., Yen, A., Sarti, G., Pal, K., Floody, O., ... & Bau, D. (2026). Agents of chaos. *arXiv preprint arXiv:2602.20021*.

363. Hammond, L., Chan, A., Clifton, J., Hoelscher-Obermaier, J., Khan, A., McLean, E., ... & Rahwan, I. (2025). Multi-agent risks from advanced ai. *arXiv preprint arXiv:2502.14143*.

364. Soares, N., Fallenstein, B., Armstrong, S., & Yudkowsky, E. (2015). *Corrigibility*. Machine Intelligence Research Institute. https://intelligence.org/files/Corrigibility.pdf.

365. Zeng, Y., Lu, E., Guo, X., Huangfu, C., Xie, J., Chen, Y., ... & Younas, A. (2025). AI Governance InternationaL Evaluation Index (AGILE Index) 2025. *arXiv preprint arXiv:2507.11546*.

366. Bommasani, R., Kapoor, S., Klyman, K., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2023, December 13). *Considerations for governing open foundation models*. Stanford Institute for Human-Centered Artificial Intelligence. https://hai.stanford.edu/policy/issue-brief-considerations-governing-open-foundation-mode

367. Tzachor, A., Devare, M., Richards, C., Pypers, P., Ghosh, A., Koo, J., ... & King, B. (2023). Large language models and agricultural extension services. *Nature food*, *4*(11), 941-948.

368. Moor, M., Banerjee, O., Abad, Z. S. H., Krumholz, H. M., Leskovec, J., Topol, E. J., & Rajpurkar, P. (2023). Foundation models for generalist medical artificial intelligence. *Nature, 616*, 259–265. https://doi.org/10.1038/s41586-023-05881-4

369. Bommasani, R., Klyman, K., Kapoor, S., Longpre, S., Ramaswami, A., Zhang, D., Schaake, M., Ho, D. E., Narayanan, A., & Liang, P. (2024). *The foundation model transparency index v1.1*. arXiv:2407.12929. https://arxiv.org/abs/2407.12929

370. BigScience Workshop, Le Scao, T., Fan, A., et al. (2022). *BLOOM: A 176B-parameter open-access multilingual language model*. arXiv. https://doi.org/10.48550/arXiv.2211.05100

371. Touvron, H., Lavril, T., Izacard, G., et al. (2023). *LLaMA: Open and efficient foundation language models*. arXiv. https://arxiv.org/abs/2302.13971

372. Qwen Team. (2024). *QwenLM*. GitHub. https://github.com/QwenLM/Qwen

373. Qwen Team. (2024). *Qwen2 technical report*. arXiv. https://arxiv.org/abs/2407.10671

374. DeepSeek-AI. (2024). *DeepSeek-V3 technical report*. arXiv. https://doi.org/10.48550/arXiv.2412.19437

375. DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning*. arXiv. https://arxiv.org/abs/2501.12948

376. Jiang, A. Q., et al. (2023). *Mistral 7B*. arXiv. https://arxiv.org/abs/2310.06825. Mistral AI.

377. Technology Innovation Institute. (2023). *The Falcon series of open language models*. arXiv. https://arxiv.org/abs/2311.16867

378. SberDevices. (2026, March). *GigaChat-3.1: Большое обновление больших моделей*. Habr. https://habr.com/ru/companies/sberbank/articles/1014146/

379. Yandex. (2025, February 25). *YandexGPT 5 — в Алисе, облаке и опенсорсе*. Habr. https://habr.com/ru/companies/yandex/articles/885218/

380. Sarvam AI. (2025). *Sarvam AI models on Hugging Face*. https://huggingface.co/sarvamai

381. SB Intuitions. (2025). *Sarashina models on Hugging Face*. https://huggingface.co/sbintuitions

382. Naver Cloud. (2024). *HyperCLOVA X technical report*. arXiv. https://arxiv.org/abs/2404.01954

383. Seger, E., Dreksler, N., Moulange, R., et al. (2023). *Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives*. Centre for the Governance of AI. https://www.governance.ai/research-paper/open-sourcing-highly-capable-foundation-models

384. Anthropic. (2025). *Disrupting the first reported AI-orchestrated cyber espionage campaign*. https://www.anthropic.com/news/disrupting-AI-espionage

385. Partnership on AI. (2023). *PAI's guidance for safe foundation model deployment*. https://partnershiponai.org/modeldeployment/

386. International Energy Agency. (2025). *Energy and AI*. International Energy Agency. https://www.iea.org/reports/energy-and-ai

Panel Científico Internacional Independiente sobre Inteligencia Artificial

Composición y mandato

El Panel Científico Internacional Independiente sobre Inteligencia Artificial se creó en el seno de las Naciones Unidas, mediante la resolución 79/325 de la Asamblea General, conforme a los compromisos asumidos en el Pacto Digital Global y el Pacto para el Futuro.

El Panel, integrado por 40 expertos independientes nombrados por la Asamblea General para un mandato de tres años por razón de su destacada experiencia en la IA y campos afines, presenta una composición equilibrada en cuanto al género e incluye a miembros de los cinco grupos regionales de Estados Miembros, procedentes de diversas disciplinas, que abarcan la IA técnica básica, la IA aplicada, la seguridad e infraestructura y las políticas, la ética y el impacto de la IA.

El Panel tiene el mandato de emitir evaluaciones científicas empíricas que sinteticen y analicen la investigación existente relacionada con las oportunidades, los riesgos y las repercusiones de la IA, mediante un informe resumido anual pertinente para las políticas, pero no prescriptivo, incluidos los resúmenes temáticos que considere necesarios. Su labor científica debe regirse por los principios de independencia, credibilidad y rigor científicos y por una participación multidisciplinaria e inclusiva.

El Panel también tiene el mandato de presentar su informe resumido anual en el Diálogo Mundial de las Naciones Unidas sobre la Gobernanza de la Inteligencia Artificial. Al contribuir al Diálogo Mundial y a otros procesos internacionales más amplios, el Panel facilita que la comunidad mundial anticipe los desafíos emergentes, adopte decisiones de gobernanza mejor fundamentadas y contribuya a equilibrar el acceso a la información entre los responsables de formular políticas de todo el mundo.

El Panel cuenta con el apoyo de la secretaría del Panel, coordinada por la Oficina de Tecnologías Digitales y Emergentes de las Naciones Unidas.

Proceso que ha llevado al presente informe preliminar

En los tres meses transcurridos desde su primera reunión, celebrada en marzo de 2026, tras su nombramiento en febrero de 2026, el Panel ha trabajado intensamente para elaborar el presente informe. Este proceso intensivo de intercambio científico y análisis colectivo incluyó una reunión plenaria presencial de tres días y más de 60 reuniones virtuales del Panel, facilitadas por sus Copresidentes elegidos.

Este informe preliminar sirve de punto de partida y sienta las bases fundamentales para una consulta más amplia con expertos externos y para la preparación de futuros resúmenes temáticos e informes anuales.

Independencia científica del Panel

Los miembros del Panel actúan a título personal en calidad de expertos científicamente independientes. Al haber sido designados como expertos de las Naciones Unidas en misión, cada uno de ellos ha declarado y prometido no solicitar ni aceptar instrucciones, con respecto al cumplimiento de sus deberes, de ningún Gobierno ni de ninguna otra fuente. El Estatuto Relativo a la Condición y los Derechos y Deberes Básicos de los Funcionarios que No Forman Parte del Personal de la Secretaría y de los Expertos en Misión (ST/SGB/2002/9) también incluye cláusulas relativas a su conducta y rendición de cuentas.

Donors

The Panel Secretariat gratefully acknowledges the financial and in-kind contributions of the following governments and partners, without whom the Panel would not have been able to carry out its responsibilities:

Government of Germany
Government of Japan
Government of Spain
Omidyar Network Fund

Panel Secretariat

Coordinator

- Amandeep Singh Gill, United Nations Under-Secretary-General for Digital and Emerging Technologies

Secretariat Coordination & Support

- Quintin Chou-Lambert, United Nations Office for Digital and Emerging Technologies (UN ODET)
- Rebakah Hayoung Woo, UN ODET
- Peppi Väänänen, UN ODET

Fundraising & Logistics

- Sebastian Frank, UN ODET
- Antonieta Loaiza, UN ODET

Communications

- Karoline Hassfurter, UN ODET
- Brian Shung Seun Lau, UN ODET
- Anamika Madhuraj, UN ODET

Editing and Drafting Support*

- Jiaee Cheong, United Nations University (UNU)
- Kevin Kohler, UNU
- Max Springer, UNU

Rapporteurs

- Wernhard Berger, United Nations Industrial Development Organization
- Jin Cui, International Telecommunication Union (ITU)
- Tim Engelhardt, Office of the United Nations High Commissioner for Human Rights (UN OHCHR)
- Andrew Morritt, United Nations Department of Peace Operations
- Prateek Sibal, United Nations Educational, Scientific and Cultural Organization (UNESCO)
- Mariagrazia Squicciarini, UNESCO
- Oleksandra Vereschak, UNESCO
- Ana Gabriela Fernandez Vergara, ITU
- Li Zhou, UN OHCHR

* To ensure scientific independence of the Panel's work, Editing and Drafting Support personnel report to the Panel on substantive content/language of written outputs, while complying with coordination requirements, timelines, and sharing of produced content as part of the Panel Secretariat. For administrative purposes, they report to United Nations University.

The following United Nations Secretariat entities also supported the Panel Secretariat in the preparation of the report: Department for General Assembly and Conference Management, Department of Global Communications, and United Nations Geospatial (Office of Information and Communications Technology).