Recuadro 5.3

IMPLICACIONES PARA LA ESTABILIDAD FINANCIERA DE LA DISRUPCIÓN TECNOLÓGICA EN EL ÁMBITO DE LA CIBERSEGURIDAD

La aceleración de la innovación tecnológica está transformando el perfil de los ciberriesgos que afectan al sistema financiero. Los avances en inteligencia artificial (IA) y en otras tecnologías como la computación cuántica podrían acelerar el desarrollo de técnicas y métodos que comprometen la seguridad de los sistemas de información y generar potencialmente un salto cualitativo en sus capacidades. Esto obligaría a revisar, al menos en parte, las estrategias de ciberseguridad actuales.

Este recuadro reúne información sobre los desarrollos más recientes en este ámbito, identifica distintas implicaciones para la estabilidad financiera y discute la capacidad de reacción del sistema bancario y de su marco regulatorio y supervisor para adaptarse y contener los riesgos operacionales.

Aplicación de la inteligencia artificial para la explotación de vulnerabilidades informáticas

En gran medida, el interés creciente en esta cuestión responde a la reciente aparición de Claude Mythos, un modelo de IA desarrollado por la empresa estadounidense Anthropic, orientado a tareas avanzadas de razonamiento y programación. Según esta empresa desarrolladora, este modelo ha demostrado capacidades excepcionales para identificar y encadenar vulnerabilidades de *software*[1].

De acuerdo con la información publicada por Anthropic en su web técnica y en la documentación técnica de Claude Mythos, esta aplicación habría sido capaz de identificar un alto número de vulnerabilidades de día cero y severidad crítica en distintos componentes de *software*, algunas de las cuales llevaban décadas presentes sin ser detectadas[2].

La novedad de la aplicación de Claude Mythos para la detección de vulnerabilidades residiría en la capacidad del modelo para descubrirlas de forma automatizada y combinarlas en cadenas de explotación, reduciendo drásticamente el tiempo entre identificación y posible uso malicioso. En términos generales, estas vulnerabilidades podrían explotarse para obtener control no autorizado de sistemas, acceder a información sensible o comprometer la integridad y disponibilidad de servicios críticos.

Los distintos elementos de *software* en los que se han identificado vulnerabilidades incluyen sistemas operativos, navegadores, programas multimedia y bibliotecas. Aunque no sería necesariamente el caso, los fallos en estas últimas podrían permitir potencialmente falsificar certificados o descifrar comunicaciones privadas.

Conviene subrayar, no obstante, que en el momento actual la información se limita a los anuncios de la empresa desarrolladora Anthropic, que las capacidades de Claude Mythos no han sido aún plenamente validadas de forma independiente y que este anuncio podría incorporar también elementos de posicionamiento comercial en un entorno de fuerte competencia entre desarrolladores de modelos avanzados de IA.

Bajo escenarios adversos, este desarrollo tecnológico alcanzaría un potencial impacto sistémico negativo. Acelerar y aumentar la escala de las capacidades de explotación de vulnerabilidades informáticas podría generar, con respecto a la situación actual, olas más sincronizadas y extensas de ciberataques que afecten a tecnologías compartidas por gran parte del sector financiero, así como por otros sectores económicos. En particular, la explotación de vulnerabilidades en elementos críticos del sistema de pagos podría dificultar o impedir durante ciertos períodos el desarrollo de transacciones económicas. Dado el uso extendido de ciertos elementos tecnológicos (por ejemplo, sistemas operativos) y la concentración de muchas empresas en un número reducido de proveedores críticos de *software* y *hardware*, este escenario de alta capacidad de la IA para el desarrollo de ciberataques elevaría la dimensión sistémica del riesgo operativo.

Conviene remarcar que el caso de Claude Mythos se enmarca en una tendencia más general de desarrollo de modelos basados en IA. En este sentido, incluso si las capacidades de este modelo concreto no resultasen finalmente excepcionales en relación con otras

1 Encadenar vulnerabilidades de *software* implica la combinación secuencial de varios fallos de seguridad distintos para construir un ataque con un impacto mucho mayor que la explotación de vulnerabilidades individuales. Por ejemplo, una vulnerabilidad permite obtener acceso inicial al sistema, otra escalar privilegios de acceso en el mismo y una tercera extraer información.

2 Una vulnerabilidad de día cero es un fallo de seguridad desconocido por el desarrollador de un *software* o *hardware* informático, sin solución disponible en el momento de ser explotado por un atacante. El término indica que los desarrolladores tienen cero días para su solución antes de que sea utilizado con fines maliciosos.

alternativas existentes, seguiría siendo un paso más en la mejora progresiva del conjunto de herramientas a disposición de usuarios legítimos, pero también de posibles actores maliciosos. En cualquier caso, la previsible aparición de más herramientas con altas capacidades facilitaría la proliferación de su uso y extendería la dimensión sistémica de este desarrollo tecnológico.

En este contexto, Anthropic ha lanzado la iniciativa *Glasswing,* que permite el acceso cerrado a Claude Mythos a un número reducido de empresas estadounidenses del sector tecnológico, de la ciberseguridad, las finanzas y el *software* abierto. Esto posibilitaría el desarrollo de estrategias defensivas de manera anticipada a la distribución comercial más amplia de esta aplicación de IA.

Esta iniciativa podría ayudar a conocer y a mitigar los riesgos asociados a esta tecnología, pero la contención efectiva de su dimensión sistémica requerirá de un grado elevado de coordinación internacional. En una economía mundial con importantes interconexiones entre empresas y entidades de distintos países, si las vulnerabilidades no se mitigan en partes extensas de la red, esto podría facilitar ataques que acabaran transmitiéndose al conjunto del sistema global. En particular, el Consejo de Estabilidad Financiera[3] (FSB, por sus siglas en inglés) podría ser uno de los foros que contribuyeran a esta coordinación internacional.

Ante el alcance sistémico y global de estas tecnologías, se plantean como necesarios tanto el acceso a *Glasswing* o a iniciativas análogas de empresas de otros países, además de las de EEUU, como la interacción regular entre sector público y privado. Ante un potencial riesgo de naturaleza sistémica, es ineludible disponer de mecanismos de análisis y testeo de las nuevas tecnologías y comparación de información globales que eviten asimetrías.

Adicionalmente, resulta importante que en el ámbito europeo se desarrollen capacidades robustas en estas tecnologías, que permitan avanzar en la soberanía digital, mejorar las capacidades propias de ciberdefensa y asegurar un rol activo en el diseño e implementación de soluciones globales frente a estos riesgos.

De forma clave, la potencial aplicación de esta tecnología para mejorar los sistemas de defensa y reducir los tiempos de reacción frente a vulnerabilidades obliga a cambiar los modelos de decisión de las entidades, haciéndolos más ágiles, y genera nuevas incertidumbres sobre la calidad técnica de las soluciones para corregir los riesgos. En conjunto, el balance entre las mejoras de las capacidades ofensivas y defensivas en ciberseguridad posibilitadas

Esquema 1
Modelo de inteligencia artificial Claude Mythos

	¿Qué es?	¿Por qué es relevante?
Claude Mythos	• Modelo de IA con capacidades avanzadas para identificar y explotar vulnerabilidades de software. • Desarrollado por la empresa estadounidense Anthropic.	• Podría cambiar el modelo de amenaza al acelerar y escalar la detección y explotación de vulnerabilidades en los sistemas de información.
Riesgo tecnológico principal	• Aumento de velocidad, escala y automatización de los ataques a sistemas de información. • Esta capacidad es actualmente incierta en ausencia de validación independiente y extensa.	• Los sistemas de defensa actuales podrían verse superados por la velocidad y las nuevas clases de ciberataques. • Podría generar oleadas más sincronizadas de explotación de vulnerabilidades, difíciles de gestionar por los sistemas actuales.
Proyecto *Glasswing*	• Proyecto de acceso limitado a Claude Mythos a distintas empresas de EEUU para desarrollar estrategias de defensa antes de su potencial distribución comercial más amplia.	• Proporcionaría ventaja a las empresas participantes para desarrollar estrategias de defensa frente a los ciberataques basados en esta tecnología.
Implicación sistémica	• Aumento de los ciber-incidentes en el conjunto del sistema económico. • Las capacidades de Claude Mythos pueden ser replicadas y expandidas por otros modelos.	• Aumento de los riesgos operacionales de todos los sectores económicos.

FUENTE: Banco de España.

3 El FSB es un organismo internacional creado en 2009 que supervisa y recomienda políticas para fortalecer el sistema financiero global. El FSB contribuye a la monitorización global de vulnerabilidades y riesgos para la estabilidad financiera, y promueve la implementación coherente de normas y la coordinación de políticas en el ámbito de la estabilidad financiera.

por la IA es hoy en día incierto. El esquema 1 resume los principales elementos de la información pública disponible sobre Claude Mythos y sus implicaciones sobre riesgos operacionales.

Computación cuántica y criptografía

El desarrollo de la computación cuántica plantea implicaciones potencialmente más profundas para la ciberseguridad que la IA, pero es una tecnología en una fase más temprana de desarrollo y cuya evolución está sujeta a más incertidumbre. Aunque su impacto práctico seguiría siendo limitado en el corto plazo, los avances en computación cuántica plantean riesgos significativos para los estándares criptográficos actuales, que sustentan la confidencialidad, integridad y autenticación de las operaciones financieras. La eventual necesidad de migrar hacia criptografía postcuántica implicaría esfuerzos técnicos y organizativos relevantes, que requieren planificación anticipada.

La computación cuántica es un paradigma de cálculo que utiliza propiedades de la mecánica cuántica (una rama de la física) para procesar información, permitiendo resolver ciertos problemas mucho más rápido que los ordenadores clásicos[4]. Los sistemas criptográficos actuales basan buena parte de su seguridad en la complejidad de determinados problemas matemáticos que son muy difíciles de resolver para los ordenadores convencionales, pero que resultan sustancialmente más sencillos para la computación cuántica. No obstante, estas tecnologías afrontan todavía retos importantes para la extensión de su aplicación práctica, como la eliminación de errores de cómputo, la coherencia de resultados, los elevados costes energéticos y las condiciones físicas extremas para su operación, así como la ausencia de soluciones estándar aplicables a escala.

A pesar de estos retos, en el mes de marzo de este año se publicaron dos artículos[5] que alertan sobre la probabilidad de que el tiempo que conlleve disponer de una máquina capaz de romper parte de la criptografía actual sea menor de lo que se creía. Estos artículos apuntan a que los recursos cuánticos necesarios para atacar ciertos esquemas de clave pública podrían ser sensiblemente menores de lo estimado hasta ahora[6]. Por tanto, aunque el riesgo sigue sin ser inmediato, pasa a ser más creíble y cercano desde una perspectiva prudencial. Ello requiere la

Esquema 2
Comparativa de desarrollo tecnológico e implicaciones sobre riesgos de la inteligencia artificial y la computación cuántica

	Inteligencia Artificial (IA)	Computación cuántica
Impacto potencial en ciberseguridad	Podría acelerar y automatizar la identificación y explotación de vulnerabilidades. Reduciría el tiempo entre descubrimiento y ataque, pero también mejora defensas.	Riesgo disruptivo para la criptografía actual. Impacto futuro e incierto, concentrado en sistemas que no migren a estándares poscuánticos.
Grado de madurez	Alta en aplicaciones generales. Difusión potencialmente rápida de modelos avanzados con capacidades emergentes en ciberseguridad.	Baja/Media. Avances relevantes en hardware y corrección de errores, pero sin aplicaciones generalizadas fuera de entornos experimentales.
Tiempos para desarrollo adicional	Corto plazo (1-3 años). Usos operativos ya observables y escalables.	Medio largo plazo. Impacto potencial elevado pero con incertidumbre en el horizonte temporal.
Probabilidad de materialización	Alta. Barreras de entrada reducidas, reutilización de infraestructuras existentes y rápida adopción por actores legítimos y maliciosos.	Media baja. Depende de hitos técnicos aún no consolidados y de inversiones sostenidas en el tiempo.

FUENTE: Banco de España.

4 Los progresos en computación cuántica se integran dentro de un conjunto más amplio de avances en tecnologías cuánticas. Otras áreas destacadas dentro de esta línea de investigación incluyen las comunicaciones y criptografía cuánticas, que permitirían potencialmente mayor seguridad en la transmisión de la información y en su cifrado, y los sensores y metrología cuántica, que permitirían una medición física más precisa de múltiples fenómenos que la permitida por las tecnologías actuales.

5 Véase Babbush et al. (2026). "Securing Elliptic Curve Cryptocurrencies against Quantum Vulnerabilities: Resource Estimates and Mitigations". arXiv:2603.28846 y Cain et al. (2026). "Shor's algorithm is possible with as few as 10,000 reconfigurable atomic qubits". arXiv:2603.28627.

6 Los esquemas de clave pública (o criptografía asimétrica) utilizan un par de claves matemáticamente relacionadas: una pública (compartida libremente para cifrar o verificar) y una privada (secreta, para descifrar o firmar).

necesidad de conocer bien los inventarios criptográficos, establecer un buen gobierno y fomentar la interrelación con los proveedores de tecnologías de ciberseguridad. El esquema 2 resume las diferencias en la valoración de riesgos operacionales ligados a la IA y a la computación cuántica.

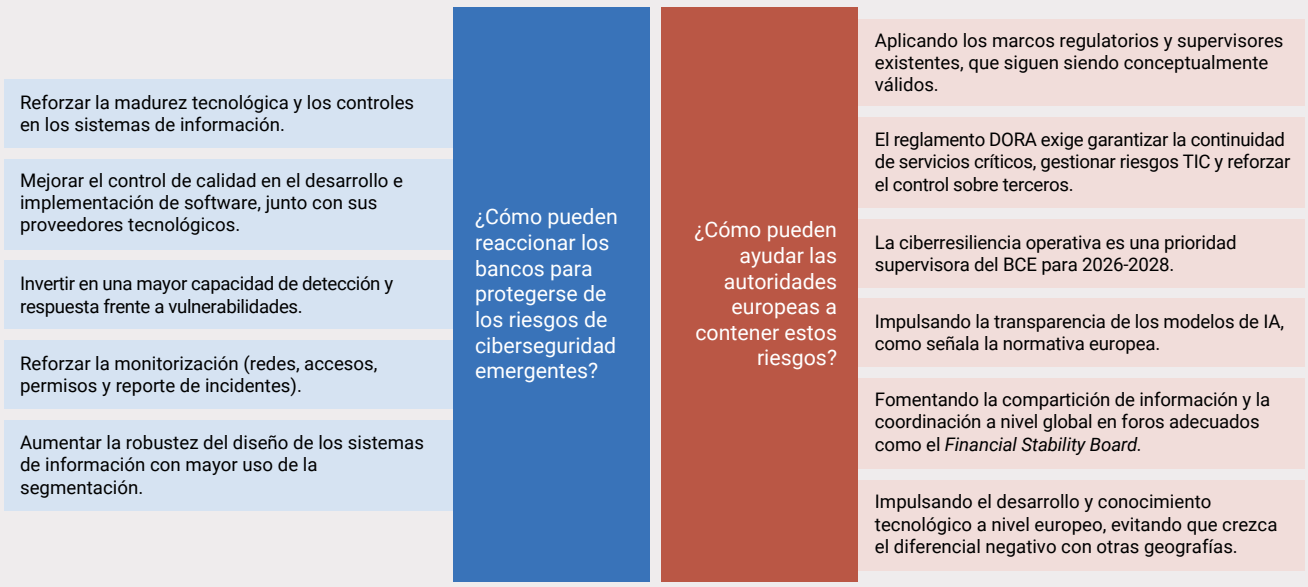
Impactos en la estabilidad financiera más allá de los efectos directos sobre riesgo operacional

Desde una perspectiva de estabilidad financiera, es también relevante el efecto sobre la confianza en el sistema financiero de este tipo de amenazas tecnológicas sobre la ciberseguridad. A nivel de entidades individuales, existe ya alguna evidencia sobre como los ciberataques pueden inducir salidas de fondos (abruptas o progresivas) hacia entidades o inversiones alternativas con una mayor percepción de solidez, incluso aunque esta no sea efectivamente superior[7].

No obstante, para que estos eventos derivasen en un episodio sistémico de volatilidad en los flujos de fondos y así de tensiones de liquidez, sería necesario que el uso indebido de estos modelos facilitara ataques simultáneos, severos y prolongados en el tiempo sobre varias entidades (o sobre proveedores tecnológicos ampliamente utilizados por el sector financiero), que indujeran desconfianza sobre un conjunto de relevancia sistémica de intermediarios o instrumentos financieros. Estos son escenarios distintos de los de ciberataques aislados, dado que estos últimos difícilmente tendrían implicaciones relevantes para la estabilidad financiera[8].

Otro canal por el cual esta nueva tecnología puede acabar afectando a la estabilidad financiera es a través de su impacto a medio plazo en la rentabilidad del sistema bancario (y también de otros intermediarios), en el caso de que se asentara un escenario de ciberincidentes severos y recurrentes. Por un lado, las entidades podrían requerir de

Esquema 3
Capacidad de reacción de los bancos y del sistema de supervisión y regulación frente a los riesgos emergentes



FUENTE: Banco de España.

7 Véanse A. Kotidis y S. Schreft. (2025). "The Propagation of Cyberattacks through the Financial System: Evidence from an Actual Event". *Journal of Finance,* 80, pp. 3313-3358, y F. Gogolin, I. Lim y F. Vallascas. (2026). "Cyberattacks on Small Banks and the Impacto on Local Banking Markets". *Journal of Money, Credit and Banking.* De próxima publicación. El primer estudio muestra que los ataques a proveedores tecnológicos externos de las entidades bancarias pueden generar disrupciones en los sistemas de pagos y tensiones de liquidez, mientras que el segundo documenta que los ciberataques se asocian con una pérdida de depósitos.

8 Para una discusión más amplia sobre las posibilidades de los ciberataques para derivar en una crisis de estabilidad financiera, véase la sección 5.4 del IEF de primavera de 2025.

mayores y más frecuentes inversiones en ciberseguridad para hacer frente a estas nuevas amenazas. Por otro lado, el coste de obtención de fondos podría encarecerse para algunos productos o contrapartes, en línea con el mencionado trasvase entre entidades o instrumentos financieros durante episodios de pérdida de confianza.

Además, por el lado del activo, un aumento sustancial y sostenido de los ciberataques podría acabar afectando a la solvencia de los sectores empresariales más expuestos, con potenciales implicaciones sobre las pérdidas por riesgo de crédito y de mercado de las entidades[9]. En resumen, a través del canal de rentabilidad, los ciberataques pueden reducir la fortaleza del sistema bancario para afrontar cualquier otro tipo de *shocks*.

Capacidad de reacción del sistema bancario y del marco regulatorio y supervisor

La mayor exposición inicial a estos riesgos podría incentivar un refuerzo adicional de la ciberresiliencia de los bancos y la adaptación del marco regulatorio y supervisor, contribuyendo así a sostener la confianza de la sociedad en el sistema bancario en su conjunto[10].

Los escenarios de riesgo planteados por estas tecnologías refuerzan la necesidad de madurez tecnológica y de controles fuertes en los sistemas de información de los bancos. Frente a estas amenazas, este sector podría, en primer lugar, reforzar el control de calidad en el desarrollo e implementación, junto con sus proveedores tecnológicos, de nuevas soluciones de *software* (por ejemplo, testeo más intenso antes de despliegue operativo), limitando las vulnerabilidades a explotar.

Adicionalmente, sería deseable invertir en aumentar la capacidad de respuesta frente a vulnerabilidades, recortando los tiempos necesarios para corregirlas y mejorando la capacidad de detectarlas. El refuerzo de la monitorización de actividades de red, de los sistemas de permisos y accesos a los sistemas de información, y del reporte de ciberincidentes[11], entre otros elementos, contribuirían a este fin. Como se ha mencionado ya, los propios avances en la tecnología de IA también pueden ser útiles para este propósito. Por ejemplo, para identificar vulnerabilidades durante el diseño de los sistemas de información para hacerlos más seguros.

Frente a estos retos tecnológicos, los marcos regulatorios y supervisores existentes siguen siendo conceptualmente válidos, y contienen los controles fundamentales para mitigar el riesgo de estos avances. Por ejemplo, el reglamento DORA *(Digital Operational Resilience Act,* por su denominación en inglés), en vigor desde enero de 2025, establece exigencias clave en materia de políticas de seguridad, pruebas periódicas y gestión de riesgos tecnológicos y frente a terceros, como los proveedores de servicios de Tecnologías de la Información y la Comunicación (TIC).

En el ámbito supervisor, el BCE ya había situado la ciber-resiliencia operativa en el centro de sus prioridades supervisoras dentro del Mecanismo Único de Supervisión para el ciclo 2026-2028[12], ampliando además la supervisión directa a proveedores TIC críticos. Este marco se complementa con pruebas avanzadas como el *Threat Led Penetration Testing* (TLPT), orientadas a evaluar la capacidad de las entidades más relevantes para detectar, responder y recuperarse ante ciberataques sofisticados, reforzando así la resiliencia operativa del sistema financiero en su conjunto.

A pesar de todas estas fortalezas, estos marcos fueron diseñados bajo supuestos distintos sobre la velocidad, automatización y escala de los atacantes. El desafío no es crear nuevos regímenes, sino actualizar supuestos, escenarios y prácticas, reforzando pruebas, planes de respuesta y mecanismos de gobernanza que permitan decisiones ágiles en contextos de elevada incertidumbre. Por ejemplo, en las pruebas TLPT se deberá garantizar

9 Véanse R. Jamilow, H. Rey y A. Tahoun. (2025). "The Anatomy of Cyber Risk". NBER Working Papers, 28906. National Bureau of Economic Research, y P. Akey, S. Lewellen, I. Liskovich y C. Schiller. (2026). "Hacking Corporate Reputations". *The Review of Finance.* De próxima publicación. Estos trabajos muestran que una mayor exposición de las empresas a ciberataques estaría asociada con una menor rentabilidad.

10 En particular, la adopción de medidas supervisoras orientadas a garantizar estándares mínimos puede favorecer las inversiones en ciberseguridad por parte de las entidades. Véanse T. Ahnert, M. Brolley, D. Cimon y R. Riordan. (2024). "Cyber Risk and Security Investment", S. Kokas, F. Vallascas y N. Rowe. (2026). "Digitalization in Banking: Evidence from the New York Cybersecurity Regulation", y K. Anand, C. Duley y P. Gai. (2026). "Cybersecurity and Financial Stability". *The Review of Finance.* De próxima publicación.

11 Los ciberincidentes son sucesos que afectan a la confidencialidad, integridad, disponibilidad o autenticidad de sistemas de información, como consecuencia de fallos técnicos, errores humanos o acciones maliciosas.

12 Véase, por ejemplo, BCE. (2026). "Upgrading banks' capacity to deal with digital risks".

que los proveedores de servicios TIC que las ejecutan tienen capacidades técnicas suficientes para emular los ciberataques más sofisticados, incluyendo aquellos que hagan uso de modelos como Claude Mythos.

Paradójicamente, a pesar de la rápida evolución de estos modelos de IA, los estándares de transparencia de los principales proveedores de estas tecnologías, como los modelos de lenguaje de gran tamaño[13], no están mejorando. El Índice de Transparencia de los Modelos Fundacionales, elaborado por investigadores de las universidades de Stanford, Princeton y MIT, muestra que, de media, los modelos de lenguaje de gran tamaño cumplen menos de la mitad de los requisitos de transparencia. La puntuación global se sitúa en 41 sobre 100 y, de forma notable, ello supone un descenso de 17 puntos con respecto a 2024, lo que devuelve los niveles de transparencia a los observados en 2023[14]. Este es un resultado claramente subóptimo que muestra la necesidad de avanzar en este terreno. La transparencia es uno de los elementos fundamentales del reglamento europeo de IA, por lo que, en este sentido, el retraso en su implementación no está ayudando a mejorar en este ámbito.

Conclusiones

Conviene evitar lecturas alarmistas sobre el potencial impacto de los ciberriesgos sobre la estabilidad financiera. No existen evidencias de una amenaza sistémica inmediata, ya sea derivada de modelos de IA como Mythos o de los avances de la computación cuántica. Sin embargo, la experiencia demuestra que los avances tecnológicos tienden a difundirse con rapidez. Esto abre una ventana limitada para la preparación frente a ciberriesgos, durante la cual resulta clave reforzar capacidades técnicas, mejorar la coordinación entre entidades, supervisores y proveedores tecnológicos, e invertir en talento especializado. En esta fase de preparación, la coordinación internacional resulta fundamental para aumentar la resiliencia a nivel global y evitar que ataques que inicialmente afecten a los agentes menos preparados se difundan a través de interconexiones.

En última instancia, en un entorno de disrupción tecnológica acelerada, la ciberresiliencia del sistema financiero dependerá especialmente de la velocidad de detección, remediación y coordinación, así como de la solidez de la gobernanza y de la anticipación de riesgos emergentes.

13 Los modelos de lenguaje de gran tamaño basados en inteligencia artificial son entrenados con grandes volúmenes de texto, que les permiten comprender y generar lenguaje natural.

14 Véase, por ejemplo, Foundational Model Transparency Index.