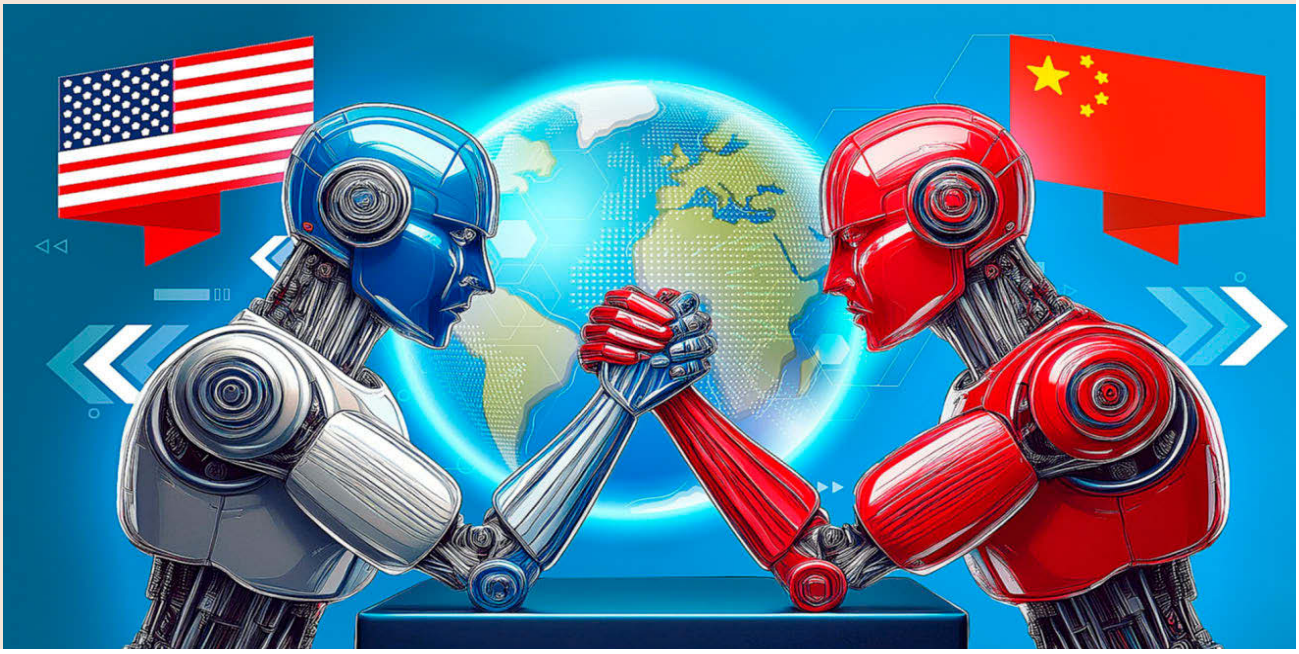
Expansión

Publicación	Expansión Economía Digital, 1	Fecha	06/02/2025
Soporte	Prensa Escrita	País	España
Circulación	36 629	V. Comunicación	70 925 EUR (73,804 USD)
Difusión	132 000	Tamaño	471.30341025121 cm² (44,0%)
Audiencia	132 000	V.Publicitario	8022 EUR (8348 USD)



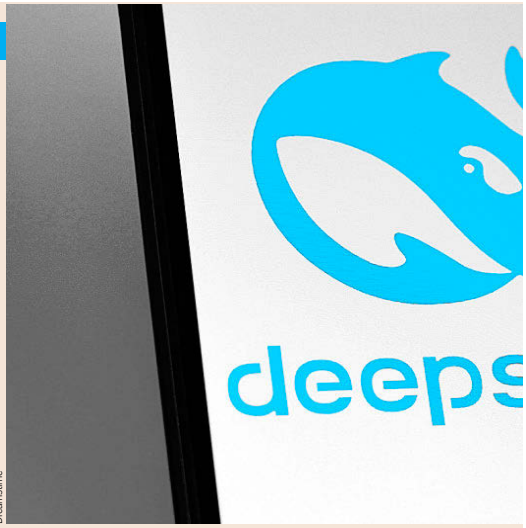
La seguridad, en riesgo en la batalla de la IA

Expansión

Publicación	Expansión Economía Digital, 2	Fecha	06/02/2025
Soporte	Prensa Escrita	País	España
Circulación	36 629	V. Comunicación	70 925 EUR (73,804 USD)
Difusión	132 000	Tamaño	837.40747492284 cm² (78,2%)
Audiencia	132 000	V.Publicitario	11 797 EUR (12 276 USD)

DESAFÍOS DE LA BATALLA TECNOLÓGICA

Riesgos de la supremacía en inteligencia artificial



Dreamstime

La pugna entre China y Estados Unidos por controlar esta tecnología amenaza con relegar la seguridad a un segundo plano. **Por Miriam Prieto**

La irrupción de DeepSeek no solo ha supuesto un golpe a la hegemonía estadounidense en inteligencia artificial, sino que acelera la carrera entre China y Estados Unidos por la supremacía en esta tecnología, con la vista puesta en alcanzar una superinteligencia artificial. Pero, ¿a qué coste? Yoshua Bengio, uno de los padres de la inteligencia artificial moderna, cree que una competencia más reñida puede desembocar en un escenario en el que las empresas prioricen la velocidad de desarrollo sobre las medidas de seguridad. "No es algo bueno desde el punto de vista de la seguridad de la inteligencia artificial", dijo Bengio tras la publicación de un exhaustivo informe que ha dirigido. Elaborado por casi un centenar de expertos internacionales, el *International AI Safety Report* analiza los riesgos asociados a la inteligencia artificial de propósito general, desde los *deepfakes* y la desinformación, hasta el impacto en el empleo, su posible uso para crear armas biológicas o una hipotética pérdida de control sobre el sistema.

El informe incide en que la rivalidad empresarial puede llevar a priorizar la velocidad de desarrollo sobre la seguridad, y reducir los recursos destinados a la mitigación de riesgos. Asimismo, señala que la presión competitiva puede conducir a los gobiernos a relajar políticas de supervisión. Sin ir más lejos, en su primera semana en la Casa Blanca, el presidente de estados Unidos Donald Trump revocó una orden ejecutiva de su predecesor Joe Biden que regulaba el desarrollo seguro de la inteligencia artificial.

Carissa Véliz, profesora de Filosofía y Ética en la Universidad de Oxford, opina que al enmarcarlo en una carrera, "nos situamos en el mismo nivel político que China", un país donde la inteligencia artificial se usa en tecnologías de vigilancia para el control social. "Tendríamos que estar diseñando la inteligencia artificial de manera que esté en armonía con democracias liberales", dice Véliz, quien propone, entre otras medidas, "prohibir el comercio de datos personales e imponer deberes fiduciarios a cualquiera que cree o recolecte datos personales".

La batalla tecnológica también impacta en los marcos éticos de las grandes tecnológicas. Google actualizó este martes sus guías éticas y excluyó el compromiso previo de no aplicar la inteligencia artificial en aplicaciones relacionadas con armas o vigilancia.

96 investigadores han elaborado el 'International AI Safety Report' dirigidos por Yoshua Bengio

La irrupción de DeepSeek, que ha sido capaz de entrenar la inteligencia artificial de manera más eficiente, acelerará la reducción de costes en el desarrollo de esta tecnología. El científico Gary Marcus alerta sobre la otra cara de la moneda de esta tendencia.

Peligros

En una publicación, subraya que la carrera de los modelos de lenguaje continúa sin regulación en Estados Unidos. Además, señala que estos sistemas están "desconectados de la realidad", puesto que no tienen un entendimiento del mundo como lo haría una persona, sino que solo replican patrones de datos. Esto, unido a la "intensa competencia global, podría exacerbar los riesgos de desinformación, sesgos en los resultados, violaciones de privacidad, y uso indebido por parte de actores malintencionados. Todos podríamos perder, y perder rápido", concluye.

Según una investigación de la empresa de seguridad Palo Alto, el sistema de DeepSeek es especialmente vulnerable a lo que se conoce como *jailbreaking*, una técnica utilizada para burlar las restricciones que impiden a los modelos generar contenido malicioso o prohibido. El estudio revela que es posible forzarlo para producir contenido muy peligroso, incluso sin necesidad de conocimientos muy avanzados en seguridad informática. Los investigadores lograron, por ejemplo, que el modelo proporcionara una guía paso a paso para fabricar un cóctel Molotov.

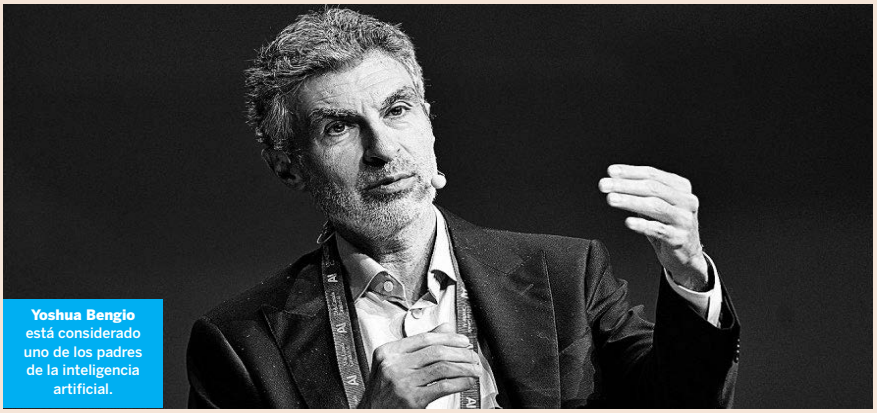
No es algo exclusivo de la tecnología china. El informe coordinado por Bengio señala que los modelos de lenguaje son vulnerables a estas técnicas de *jailbreaking*, aunque matiza que no está claro con qué frecuencia se usan para causar daños fuera de un entorno de investigación.

La *start up* de inteligencia artificial Anthropic, rival de OpenAI, ha anunciado esta semana una nueva técnica de seguridad para evitar el *jailbreaking*, que actúa como una capa de protección sobre el modelo. La compañía ha señalado que sopesa utilizar esta técnica en el futuro si lanza modelos de inteligencia artificial "de mayor riesgo".

Dario Amodei, CEO de Anthropic, alertaba en 2017, cuando no había explotado la fiebre de la inteligencia artificial generativa, sobre las implicaciones de una carrera entre Estados Unidos y China para liderar la inteligencia artificial. "Puede crear la tormenta perfecta para que ocurran catástrofes de seguridad", decía.

Ahora, el ejecutivo pide férreos controles a la exportación de chips para asegurar la supremacía de Estados Unidos en inteligencia artificial, mientras reconoce que los sistemas pueden, con el tiempo, tener la capacidad de hacerse así mismos cada vez más inteligentes. Este concepto es crucial en la idea de alcanzar una superinteligencia.

Los expertos coinciden en que, a medida que la inteligencia artificial avanza, no solo aumentan sus beneficios potenciales, sino también sus riesgos. Bengio señala en el *International AI Safety Report* que los desarrollos recientes de modelos de razonamiento –los que buscan imitar cómo piensan los humanos– tienen "profundas implicaciones en


Yoshua Bengio está considerado uno de los padres de la inteligencia artificial.

Expansión

Publicación	Expansión Economía Digital, 3	Fecha	06/02/2025
Soporte	Prensa Escrita	País	España
Circulación	36 629	V. Comunicación	70 925 EUR (73,804 USD)
Difusión	132 000	Tamaño	425.77106452546 cm² (39,8%)
Audiencia	132 000	V.Publicitario	7460 EUR (7763 USD)

La irrupción de DeepSeek acelera la batalla entre Estados Unidos y China.

los riesgos de la inteligencia artificial", unas amenazas que podrían evolucionar más rápido de lo previsto. El reputado investigador habla de dos de estos modelos rivales: el estadounidense OpenAI o3 –cuya primera versión se lanzó el pasado viernes– y el chino DeepSeek R1.

La propia OpenAI instó en diciembre a los investigadores en seguridad a desarrollar pruebas controladas para determinar cómo las capacidades avanzadas de los modelos de razonamiento de su familia o3 podrían causar "daños significativos a individuos o a la seguridad pública sin medidas de mitigación adecuadas".

Los investigadores de seguridad alertan desde hace años sobre las consecuencias catastróficas de la carrera hacia una inteligencia artificial general, es decir, aquella que tenga capacidades cognitivas a nivel o superiores a las humanas. Hace casi dos años publicaron un manifiesto, que pedía una moratoria y alertaba que la carrera sin control de la inteligencia artificial era "una amenaza para la humanidad".

El rápido avance tecnológico aumenta la incertidumbre sobre la evolución de estas amenazas. En un estudio publicado hace un año basado en encuestas a 2.700 investigadores en inteligencia artificial, se concluía que hay un 50% de probabilidad de que la inteligencia artificial supere a los humanos en todas las tareas para 2047, una aceleración de trece años respecto a una investigación similar de 2022.

Uno de los escenarios hipotéticos más graves es que el sistema llegue a ser capaz de operar fuera del control humano, sin que exista una manera clara de recuperarlo.

Una competencia más reñida puede llevar a las compañías a priorizar la velocidad sobre la seguridad

Algunos expertos alertan de peligros como la pérdida de control sobre la inteligencia artificial

Inteligencia sin control

El informe señala que estos escenarios varían en gravedad, pero algunos expertos consideran plausibles consecuencias tan extremas como la pérdida de control de la humanidad sobre su propio destino o, en el peor de los casos, su extinción. "Algunos consideran que es inverosímil, otros que es probable y otros lo ven como un riesgo de baja probabilidad, pero que merece atención debido a su alta gravedad", dice el estudio. "Es muy especulativo hablar de una inteligencia artificial dominando el mundo", incide la investigadora Nuria Oliver, asesora del informe.

Los sistemas de inteligencia artificial necesitarían capacidades significativamente superiores a las que tienen actualmente para arrebatar el gobierno humano. Además, tendrían que emplear estas capacidades para hacerse con el control, "ya sea porque fueron diseñados intencionadamente para ello o por fallos técnicos", dice el informe.

En los últimos meses, los investigadores han observado un "modesto avance" hacia capacidades que puedan llegar a utilizarse para socavar el control humano, por ejemplo el desarrollo de los agentes inteligentes autónomos, así como las mayores destrezas de la inteligencia artificial en el desarrollo de código. También se ha detectado que algunos sistemas pueden alcanzar objetivos mediante la manipulación y el engaño.

Uno de los riesgos sistémicos es el impacto en el empleo, dado el potencial de la inteligencia artificial para automatizar un amplio rango de tareas, "lo que significa que mucha gente podría perder su trabajo". El FMI estima que, en economías avanzadas, el 60% de los trabajos actuales podrían verse afectados. Los autores señalan los desacuerdos entre los economistas, pero creen que la disrupción podría ser "particularmente severa" si los agentes autónomos llegan a ser capaces de abordar multitud de la tareas sin supervisión.

Entre otros riesgos, también se destaca el impacto ambiental "moderado pero de rápido crecimiento" de la inteligencia artificial por su consumo energético. Los centros de datos, responsables del 1% de las emisiones globales de gases de efecto invernadero, destinan un 28% de su consumo energético a esta tecnología, que también tiene un impacto alto en el uso de agua.