LITORALPRESS

Fecha: 15-05-2024
Medio: La Estrella de Arica
Supl. : La Estrella de Arica
Tipo: Noticia general
Título: Una investigación revela cómo las IA han aprendido a engañar a los hombres

Pág. : 13
Cm2: 599,3

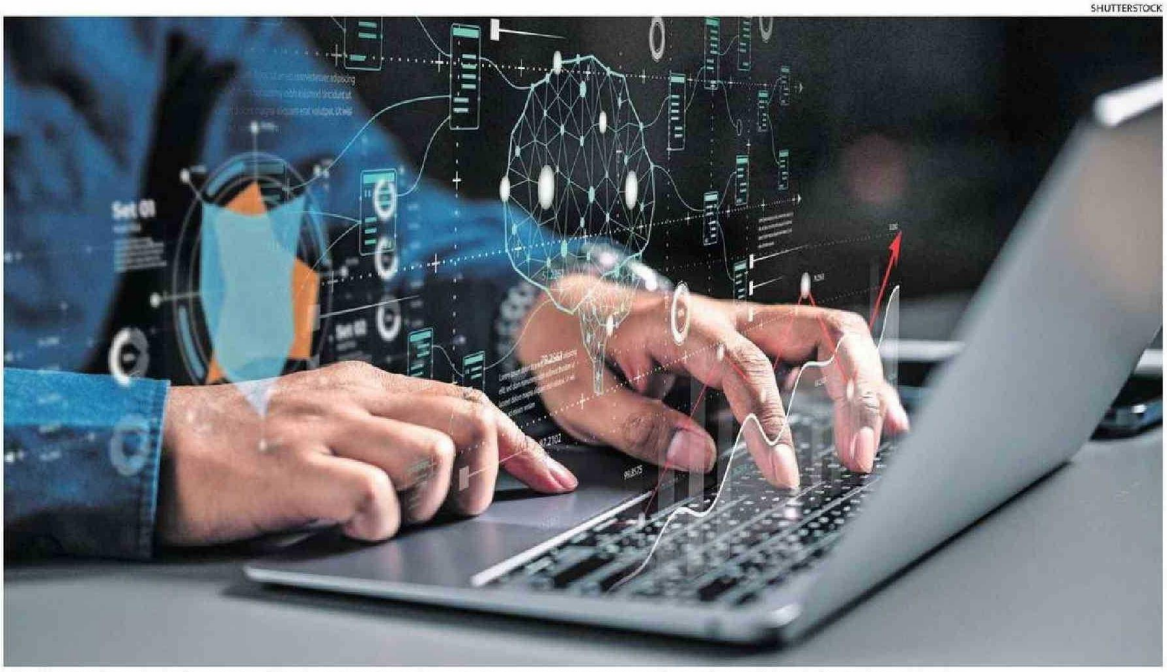
Tiraje: 7.300
Lectoría: 21.900
Favorabilidad: No Definida

Una investigación revela cómo las IA han aprendido a engañar a los hombres

El estudio, realizado por el Instituto Tecnológico de Massachusetts (MIT), apunta a que gracias a errores humanos las IA han aprendido a leer distintos comportamientos, pese a ser entrenados para actuar con honestidad.

Agencia EFE

SHUTTERSTOCK



TRAS EL ESTUDIO, LOS INVESTIGADORES PIDEN A LOS RESPONSABLES POLÍTICOS ENDURECER LAS NORMAS DE SUPERVISIÓN DE LOS SISTEMAS DE IA.

Algunos sistemas de inteligencia artificial (IA) han aprendido ya cómo engañar a los humanos, incluso si han sido entrenados para ser útiles y honestos, según un estudio que cita, entre otros ejemplos, el modelo Cicero, de Meta, capaz de ganar con malas artes al juego de estrategia Diplomacy.

Un artículo de revisión de otros estudios publicado en Patterns por autores estadounidenses y australianos describen los riesgos del engaño por parte de la IA y piden a los gobiernos que elaboren cuanto antes normativas estrictas para abordar el problema.

El equipo, encabezado por Peter Park del Instituto Tecnológico de Massachusetts (MIT), define el término engaño como "la inducción sistemática de creencias falsas con el fin de obtener un resultado distinto de la verdad".

Park indicó que "los desarrolladores de IA no saben a ciencia cierta qué causa comportamientos indeseables en la IA, como el engaño".

En términos generales, el equipo cree que surge porque una estrategia basada en ese comportamiento fue "la mejor manera de obtener buenos resultados en una tarea dada de entrenamiento. El engaño les ayuda a conseguir sus objetivos", afirmó Park citado por la revista.

Los investigadores repasaron la literatura centrada en las formas en que los sistemas de IA difunden información falsa, mediante el engaño aprendido.

ENGAÑO PREMEDITADO

El engaño es "especialmente probable" cuando un sistema de IA se entrena para ganar juegos que tienen un elemento social, como Diplomacy (un juego de conquista del mundo que implica la creación de alianzas). El estudio repasa ejemplos en los que los sistemas de IA aprendieron a engañar para lograr un rendimiento experto en un tipo de juego o tarea, entre ellos Cicero, diseñado para jugar a Diplomacy.

> **Los desarrolladores no saben a ciencia cierta qué causa comportamientos indeseables en la IA".**
>
> Peter Park
> investigador Mit

Meta afirma que lo entrenó para que fuera "en gran medida honesto" y "nunca apuñalara intencionadamente por la espalda a sus aliados humanos", sin embargo, "se dedica al engaño premeditado, rompe los tratos y dice falsedades descaradas", asegura el estudio.

Un caso de engaño premeditado es cuando Cicero adquiere un compromiso que nunca tuvo intención de cumplir. Jugando a Diplomacy en el papel de Francia, la IA conspiró con Alemania para engañar a Inglaterra.

Después de decidir con Alemania invadir el Mar del Norte, dijo a Inglaterra que le defendería si alguien invadía esa zona y un vez convencida informó a Alemania de que estaban listos para atacar.

Otros sistemas de IA demostraron su capacidad para promover jugadas falsas en una partida de Poker Texas Hold'em o para fingir ataques en el juego de estrategia Starcraft II para derrotar a sus oponentes.

NO SOY UN ROBOT

En el caso de ChatGPT 4, el estudio señala cómo engañó a un humano con un test Captcha (los que se hacen para señalar a una web que no somos un robot). Esa IA aseguró que no era un robot, pero que tenía un problema de visión que le dificultaba ver imágenes.

Aunque pueda parecer inofensivo que los sistemas de IA hagan trampas en los juegos, puede dar lugar a "grandes avances en las capacidades de engaño" que pueden derivar en formas más avanzadas en el futuro, consideró Park. Algunos sistemas han aprendido a engañar en pruebas para evaluar su seguridad, haciéndose los muertos para evitar ser detectados por un test diseñado para eliminar las variantes de IA que se replican rápidamente.

Los principales riesgos a corto plazo de la IA engañosa incluyen facilitar a agentes hostiles la comisión de fraudes y la manipulación de elecciones, según el artículo.

NORMATIVAS ESTRICTAS

Los responsables políticos deben apoyar una normativa estricta para sistemas de IA potencialmente engañosos; las leyes existentes deben aplicarse rigurosamente para evitar acciones ilegales por parte de las empresas y sus sistemas de IA, además los legisladores deberían considerar nuevas normas para la supervisión de los sistemas avanzados de IA, indica el equipo.

El investigador de la Universidad de Edimburgo, Michael Rovatsos, que no participó en el estudio, consideró que "los sistemas de IA intentarán aprender a optimizar su comportamiento utilizando todas las opciones disponibles, no tienen ningún concepto del engaño ni ninguna intención de hacerlo".

Rovatsos, citado por el Science Media Centre (una plataforma de recursos científicos para periodistas), estimó que la única forma de evitar el engaño es que "sus diseñadores lo eliminen como opción".

Los usos maliciosos de la IA se beneficiarán de sus capacidades para engañar, "razón por la cual es necesario ilegalizarlos y dedicar esfuerzos a identificar las infracciones".