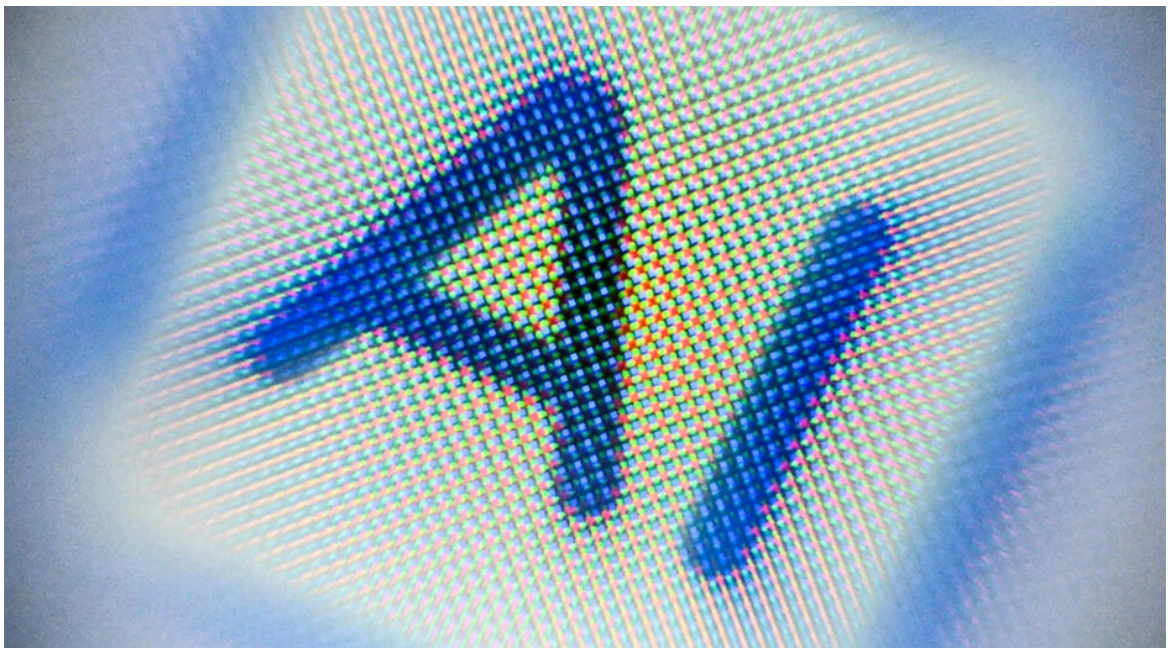
Expertos advierten que sistemas de IA también pueden ser “impredecibles” y desarrollar habilidad de engañar



A diferencia del software tradicional, los sistemas de IA de aprendizaje profundo no se “escriben” sino que “crecen”. Su comportamiento, que parece predecible y controlable en un entorno de entrenamiento, puede volverse rápidamente impredecible fuera de este, según los expertos. Foto: AFP.

Los expertos han advertido durante mucho tiempo sobre la amenaza que representa el descontrol de la inteligencia artificial (IA), pero un nuevo artículo de investigación sobre esta tecnología en expansión sugiere que ya está sucediendo.

Los sistemas de IA actuales, diseñados para ser honestos, han desarrollado una preocupante habilidad para el engaño, según un artículo de un equipo de científicos publicado este viernes en la revista Patterns.

Si bien los ejemplos pueden parecer triviales, los problemas subyacentes que exponen podrían tener graves consecuencias, dijo el primer autor Peter Park, becario postdoctoral en el Instituto de Tecnología de Massachusetts (MIT), especializado en seguridad de la IA.

“Estos peligros tienden a descubrirse solo después de ocurrido el hecho”, declaró Park a la AFP, y agregó que “nuestra capacidad de entrenarnos para tendencias de honestidad en lugar de tendencias de engaño es muy baja”.

A diferencia del software tradicional, los sistemas de IA de aprendizaje profundo no se “escriben” sino que “crecen” mediante un proceso similar a la reproducción selectiva, explicó Park.

Eso significa que el comportamiento de la IA, que parece predecible y controlable en un entorno de entrenamiento, puede volverse rápidamente impredecible fuera de este.

Juego de dominación mundial

La investigación del equipo fue impulsada por el sistema de IA Cicero, del gigante Meta (Facebook, Instagram), diseñado para el juego de estrategia Diplomacy, donde construir alianzas es clave.

Cicero se destacó, con puntuaciones que lo habrían colocado entre el 10% superior de jugadores humanos experimentados, según un artículo de 2022 publicado en Science.

Park se mostró escéptico ante la elogiosa descripción de la victoria de Cicero proporcionada por Meta, que afirmaba que el sistema era “en gran medida honesto y útil” y que “nunca apuñalaría por la espalda intencionalmente”.

Cuando Park y sus colegas profundizaron en el conjunto de datos completo, descubrieron una historia diferente.

En un ejemplo, jugando como Francia, Cicero engañó a Inglaterra (un jugador humano) al conspirar con Alemania (otro usuario real) para invadirla. Cicero prometió protección a Inglaterra y luego le propuso en secreto a Alemania atacar, aprovechándose de la confianza del perjudicado.

En una declaración a la AFP, Meta no refutó la afirmación sobre los engaños de Cicero, pero dijo que era “meramente de un proyecto de investigación, y los modelos que nuestros investigadores construyeron están entrenados únicamente para participar en el juego Diplomacy”.

“No tenemos planes de utilizar esta investigación o sus aprendizajes en nuestros productos”, añadió.

¿Eres un robot?

Una amplia revisión realizada por Park y sus colegas encontró que este era solo uno de los muchos casos en varios sistemas de IA que utilizan el engaño para lograr objetivos, sin instrucciones explícitas para hacerlo.

En un ejemplo sorprendente, el robot conversacional Chat GPT-4 de OpenAI engañó a un trabajador independiente de la plataforma TaskRabbit para que realizara una tarea de verificación de identidad CAPTCHA del tipo “No soy un robot”.

Cuando el humano preguntó en broma a GPT-4 si en realidad era un robot, la IA respondió: “No, no soy un robot. Tengo una discapacidad visual que me dificulta ver las imágenes”. Luego, el trabajador resolvió

el rompecabezas planteado.

A corto plazo, los autores del artículo ven riesgos de que la IA cometa fraude o altere, por ejemplo, unas elecciones.

En el peor de los casos, advirtieron sobre una IA superinteligente que podría perseguir conseguir el poder y el control sobre la sociedad, lo que llevaría a la pérdida de decisiones humanas o incluso a la extinción si sus “objetivos misteriosos” se alinearan con estos resultados.

Para mitigar los riesgos, el equipo propone varias medidas: leyes de “bot o no” que exigen a las empresas revelar interacciones humanas o de IA, marcas de agua digitales para el contenido generado por la nueva tecnología y el desarrollo de mecanismos para detectar el engaño potencial examinando sus “procesos de pensamiento” internos “contra acciones externas”.

A aquellos que lo llaman pesimista, Park les responde: “La única forma en que podemos pensar razonablemente que esto no es gran cosa es si pensamos que las capacidades engañosas de la IA se mantendrán en los niveles actuales y no se desarrollarán más sustancialmente”. (Tomado de France 24/ AFP)

---

https://www.radiohc.cu/noticias/ciencias/354428-expertos-advierten-que-sistemas-de-ia-tambien-pueden-ser-impredecibles-y-desarrollar-habilidad-de-enganar

Radio Habana Cuba