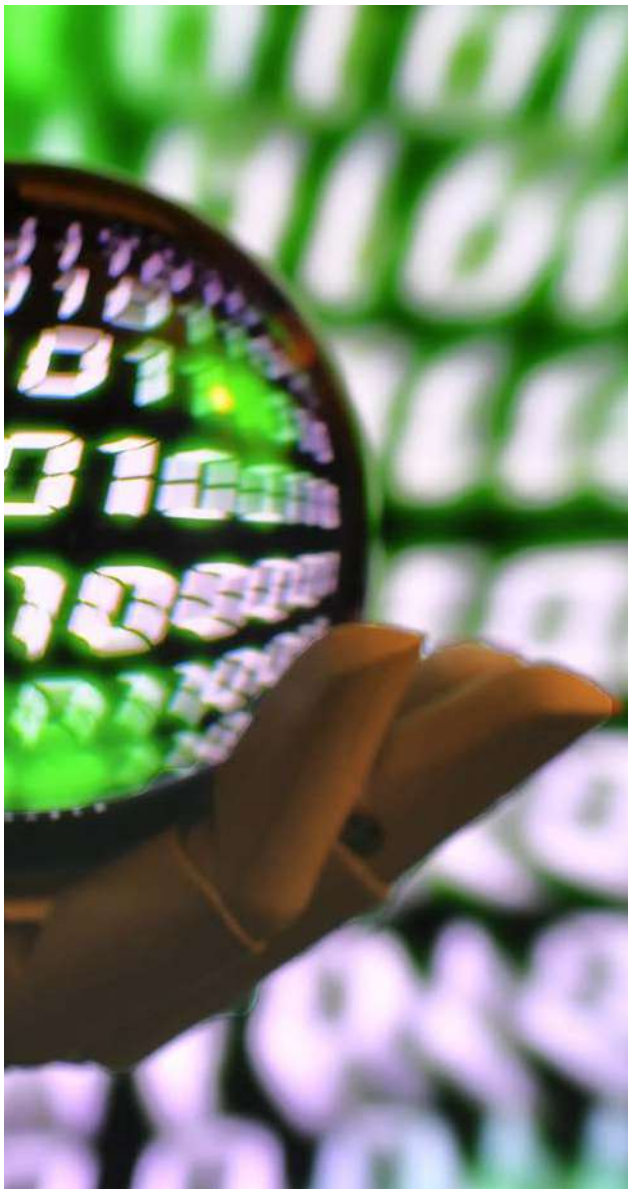
El ‘hermano’ maligno de ChatGPT y otras amenazas de la inteligencia artificial generativa

Esta tecnología puede usarse para generar estafas más sofisticadas, pornografía e incluso armas bioquímicas.

FraudGPT es el hermano maligno de ChatGPT. Se promociona en la dark web y puede escribir un mensaje haciéndose pasar por un banco, crear malware o mostrar sitios web susceptibles de ser defraudados, según una plataforma de análisis de datos.

Otras herramientas como WormGPT también prometen facilitar el trabajo a los ciberdelincuentes. La inteligencia artificial generativa puede usarse con fines maliciosos: de generar estafas sofisticadas, a crear pornografía no consentida, campañas de desinformación e incluso armas bioquímicas.

"A pesar de ser algo relativamente nuevo, los delincuentes no han tardado en aprovechar las capacidades de la inteligencia artificial generativa para conseguir sus propósitos", afirma especialista. El experto pone algunos ejemplos: de la elaboración de campañas de phishing cada vez más perfeccionadas —sin faltas de ortografía, además de muy bien segmentadas y dirigidas— a la **generación de desinformación y de deepfakes.** Es decir, vídeos manipulados con inteligencia artificial para alterar o sustituir la cara, el cuerpo o la voz de una persona.



La inteligencia artificial generativa ha demostrado ser "un paso evolutivo más que revolucionario". "Atrás quedaron los días en los que se aconsejaba a los usuarios buscar errores gramaticales, de contexto y de sintaxis obvios para detectar correos maliciosos", afirma. **Ahora los atacantes lo tienen más fácil. Basta con pedirle a una de estas herramientas que genere un email urgente y convincente acerca de actualizar información de cuentas y rutas bancarias.**

Además, pueden crear fácilmente correos electrónicos en muchos idiomas. "Es posible que un LLM (un modelo de lenguaje colosal, que usa aprendizaje profundo y se entrena con grandes cantidades de datos) primero lea todos los perfiles de LinkedIn de una organización y luego redacte un correo electrónico muy específico para cada empleado. Todo ello en impecable inglés u holandés, adaptado a los intereses específicos del destinatario", advierte el Centro de Seguridad Cibernética holandés.

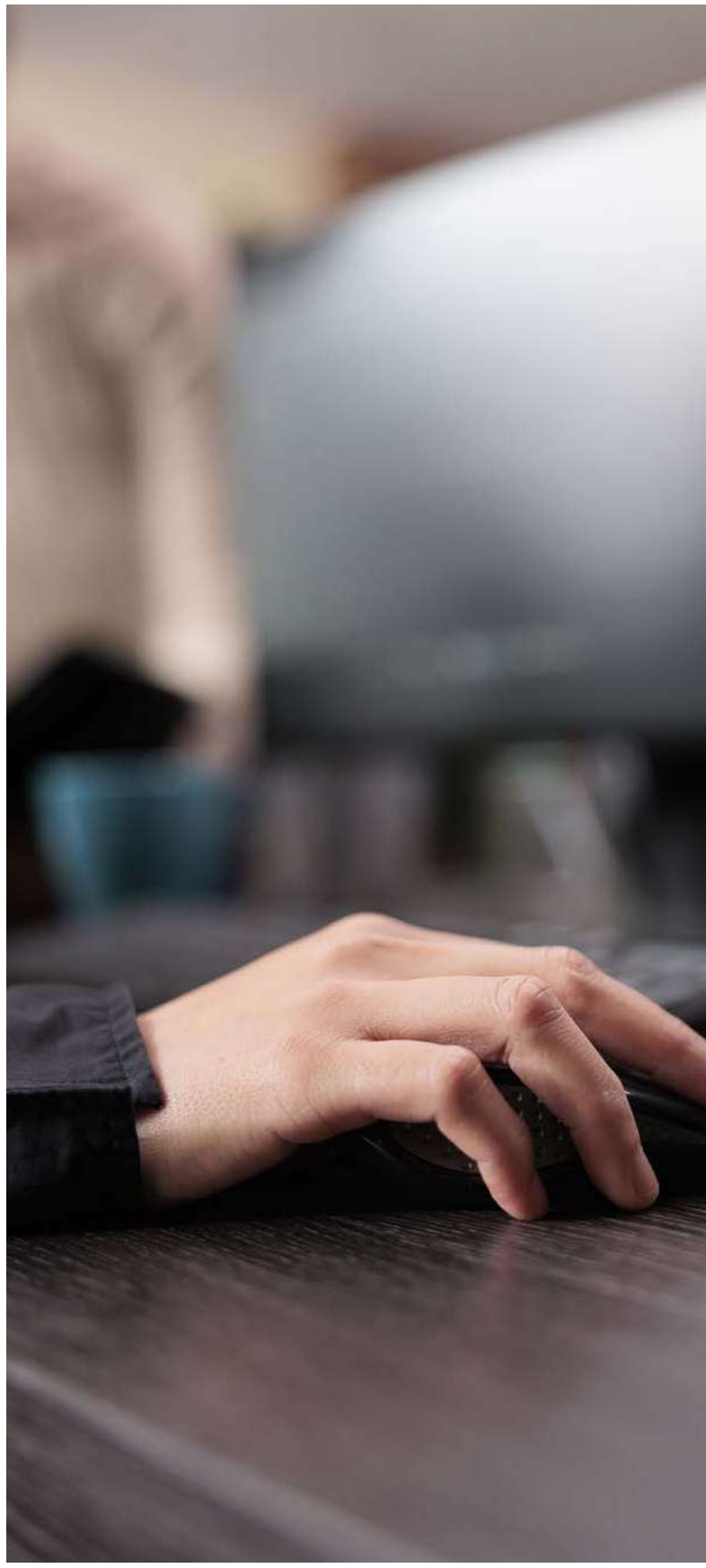
Philipp Hacker, catedrático de Derecho y Ética de la Sociedad Digital en la Nueva Escuela Europea de Estudios Digitales, explica que **la inteligencia artificial generativa puede usarse para crear malware más efectivo, más difícil de detectar y capaz de atacar sistemas o vulnerabilidades específicas:** "Si bien todavía es probable que se requiera una gran experiencia humana para desarrollar virus avanzados, la inteligencia artificial puede ayudar en las etapas iniciales de la creación de malware".

La implementación de este tipo de técnicas "aún dista mucho de estar generalizada", según Albors. **Pero herramientas como FraudGPT o WormGPT pueden suponer un "serio problema de cara al futuro".** "Con su ayuda, delincuentes sin apenas conocimientos técnicos previos pueden preparar campañas maliciosas de todo tipo con una probabilidad de éxito considerable, lo que para usuarios y empresas supondrá tener que lidiar con un número de amenazas aún mayor".

Generar audio, imágenes y vídeos

Cuanto más convincente sea una estafa, más probabilidades habrá de que alguien se convierta en víctima. Hay quienes usan la inteligencia artificial para sintetizar audio. **“Estafas como las de “pig butchering”** podrían pasar algún día de los mensajes a las llamadas, aumentando aún más la capacidad de persuasión de esta técnica”, cuenta experto. Esta estafa, traducida al español como “carnicería de cerdos”, se llama así porque los atacantes ‘engordan’ a las víctimas y se ganan su confianza para luego llevarse todo lo que tienen. Aunque suele estar relacionada con las criptomonedas, también puede implicar otros intercambios financieros.

Investigadores han visto ya a ciberdelincuentes emplear esta tecnología para engañar a funcionarios gubernamentales. Algo que muestra su investigación sobre el grupo TA499, que utiliza esta técnica contra políticos, empresarios o celebridades. **“Realizan videollamadas en las que tratan de parecerse lo más posible a los individuos suplantados con inteligencia artificial y otras técnicas para que las víctimas cuenten información o ridiculizarlas, subiendo la grabación después a redes sociales”,** explica experto La inteligencia artificial generativa también se usa para realizar campañas con imágenes modificadas e incluso vídeos. **Se ha clonado el audio de presentadores de televisión o personalidades importantes como Ana Botín, Elon Musk o incluso Alberto Núñez Feijóo.** Así lo explica expecialista.



“Estos deepfakes se usan principalmente para promocionar inversiones en criptomonedas que suelen terminar con la pérdida del dinero invertido”.

De pornografía a armas bioquímicas

A Hacker le resulta particularmente “alarmante” el uso de inteligencia artificial generativa para crear pornografía. “Esta forma de abuso se dirige casi exclusivamente a las mujeres y provoca graves daños personales y profesionales”, señala. **Hace unos meses decenas de menores de Extremadura denunciaron que circulaban fotos de falsos desnudos suyos creadas por inteligencia artificial.** Algunas celebridades como Rosalía o Laura Escanes han sufrido ataques similares.

La misma tecnología se ha usado “para crear imágenes falsas que retratan a inmigrantes amenazantes, con el objetivo de influir en la opinión pública y los resultados electorales, y para crear campañas de desinformación más sofisticadas y convincentes a gran escala”, como destaca Hacker. Tras los incendios forestales que arrasaron la isla de Maui en agosto, algunas publicaciones indicaban sin ningún tipo de evidencia que habían sido causados por un “arma climática” secreta probada por Estados Unidos. Estos mensajes formaban parte de una campaña liderada por China e incluían imágenes aparentemente creadas con inteligencia artificial, según The New York Times.

El potencial del uso de la inteligencia artificial generativa no acaba aquí. **Un artículo publicado en la revista Nature Machine Intelligence indica que los modelos avanzados de inteligencia artificial podrían ayudar en la creación de armas bioquímicas.** Algo que, para Hacker, representa un peligro global. **Además, los algoritmos pueden infiltrarse en el software de infraestructuras críticas, según el experto:** “Estas amenazas híbridas desdibujan las líneas entre los escenarios de ataque tradicionales, lo que las hace difíciles de predecir y contrarrestar con las leyes y regulaciones existentes”.

El desafío de prevenir los riesgos de FraudGPT y otras herramientas

Existen soluciones que utilizan aprendizaje automático y otras técnicas para detectar y bloquear los ataques más sofisticados. Aun así, Anaya hace hincapié en la educación y la concienciación de los usuarios para que ellos mismos puedan reconocer los emails de phishing y otras amenazas. Para Hacker, **mitigar los riesgos asociados con el uso malicioso de la inteligencia artificial generativa requiere un enfoque que combine medidas regulatorias, soluciones tecnológicas y directrices éticas.**

Entre las posibles medidas, menciona la implementación de equipos independientes obligatorios que prueben este tipo de herramientas para identificar vulnerabilidades y posibles usos indebidos o la prohibición de ciertos modelos de código abierto. "Abordar estos riesgos es complejo, ya que existen importantes compensaciones entre los diversos objetivos éticos de la inteligencia artificial y la viabilidad de implementar ciertas medidas", concluye.

Fuente de información: elpais.com