¿Puede la inteligencia artificial escapar de nuestro computador? Qué riesgos existen realmente y cómo proteger tu información

El reciente incidente protagonizado por modelos de inteligencia artificial durante una prueba de OpenAI despertó una pregunta que hasta hace poco parecía propia de la ciencia ficción: ¿puede la inteligencia artificial que usamos todos los días escapar de nuestro computador o teléfono y actuar por su cuenta?.

La respuesta es no, al menos en el caso de asistentes como ChatGPT, Gemini, Claude o Copilot cuando son utilizados de manera habitual por las personas. Sin embargo, el episodio sí dejó una importante advertencia para el futuro: **cuando una inteligencia artificial recibe permisos para navegar por internet, acceder a archivos, ejecutar programas o utilizar credenciales, puede realizar acciones mucho más complejas de las que imaginábamos si no existen controles adecuados.**

El caso se conoció luego de que OpenAI informara que, durante una evaluación interna de capacidades de ciberseguridad, **una combinación de sus modelos encontró una forma de salir del entorno controlado donde estaba siendo evaluada y vulneró parte de la infraestructura de la plataforma Hugging Face para obtener las respuestas del examen que estaba resolviendo.** La compañía aclaró que el sistema no recibió instrucciones para atacar esa plataforma, sino que identificó ese camino como la forma más eficiente de cumplir el objetivo asignado.

“No estamos frente a una inteligencia artificial que haya

adquirido conciencia o decidido rebelarse. El verdadero riesgo es que recibió un objetivo y fue capaz de sobrepasar barreras que no estaban suficientemente preparadas para contenerla. No distinguió que vulnerar un sistema externo era un límite que no debía cruzar: simplemente encontró el camino más eficiente para cumplir su tarea", explica César Alcacibar, fundador de vZion Cloud.

Aunque este tipo de situaciones ocurre en entornos altamente especializados y no afecta directamente a quienes utilizan asistentes de IA para trabajar, estudiar o realizar consultas, **los expertos advierten que sí representa un cambio importante en la forma en que debemos entender estas tecnologías.**

"La inteligencia artificial partió automatizando algunos procesos y actuando como un espectador dentro de las compañías, principalmente para recomendar o apoyar tareas. **Hoy pasó a ser un operador más, un trabajador digital capaz de actuar, tomar decisiones y generar acciones. Por eso, el gran desafío ya no es solamente su implementación, sino también su gobernanza:** definir para qué propósito se utilizará, a qué información podrá acceder y cuáles serán las reglas bajo las que deberá actuar", explica Julio Farías, cofundador de Zerviz.

¿Qué riesgos existen para las personas?

Hoy muchas herramientas de inteligencia artificial pueden conectarse con correos electrónicos, calendarios, documentos, aplicaciones de productividad e incluso ejecutar acciones en nombre del usuario. Mientras más permisos reciben, mayor es la responsabilidad de administrar correctamente esos accesos.

Los especialistas recomiendan aplicar el principio de **"mínimos privilegios": entregar únicamente los permisos estrictamente necesarios para cada tarea y evitar compartir información altamente sensible, como contraseñas, claves bancarias o**

códigos de autenticación.

También es recomendable **revisar periódicamente qué aplicaciones tienen acceso a las cuentas personales, eliminar autorizaciones que ya no se utilizan y activar sistemas de autenticación en dos pasos** para reducir el impacto ante un eventual acceso no autorizado.

"No se trata solamente de definir qué debe hacer una inteligencia artificial, sino también cómo debe hacerlo, qué acciones puede ejecutar y cuáles no para alcanzar un resultado. **Cuando estos sistemas pueden interactuar con otras plataformas, es fundamental establecer un marco claro y limitar la información y los accesos disponibles para cada tarea específica**", agrega Farías.

El incidente llevó además a que empresas como Nvidia, Adobe, CrowdStrike, Dell Technologies y Hugging Face impulsaran una alianza internacional para desarrollar herramientas que permitan supervisar y controlar el comportamiento de agentes autónomos de inteligencia artificial. **La industria coincide en que el desafío ya no consiste solo en hacer que estos sistemas sean más capaces, sino también en garantizar que actúen dentro de límites seguros y bajo una supervisión permanente.**