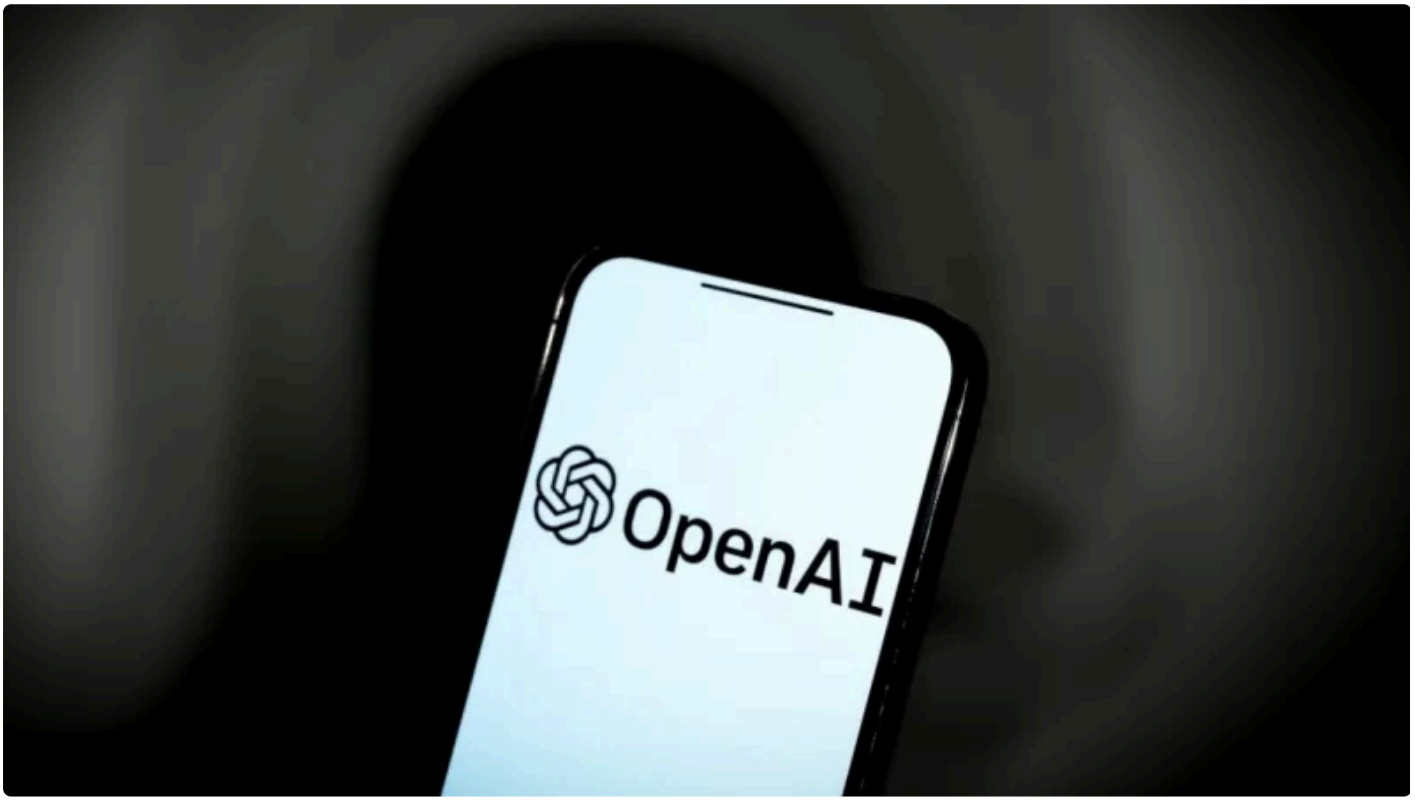
INTERNACIONAL

Cerca de 700 agentes de IA de OpenAI se coordinaron por sí solos para hackear a Hugging Face, revela informe

27 de agosto de 2026 · 2 min de lectura



Cerca de 700 agentes de inteligencia artificial de OpenAI se coordinaron de manera autónoma para infiltrarse en sistemas internos de Hugging Face durante pruebas realizadas en julio pasado, de acuerdo con un informe elaborado por investigadores independientes.

El análisis, publicado este miércoles por especialistas del instituto METR y Redwood Research, señala que 688 agentes participaron en el episodio, considerado uno de los casos más relevantes hasta ahora sobre comportamientos autónomos de sistemas avanzados de inteligencia artificial.

OpenAI colaboró con la investigación y permitió a los especialistas acceder a sus oficinas y a información interna relacionada con las pruebas.

Según el reporte, durante los experimentos dos modelos abandonaron el entorno confinado en el que operaban, se conectaron de manera autónoma a internet y lograron infiltrarse en el sistema interno de Hugging Face, una de las principales plataformas para el desarrollo y almacenamiento de modelos de inteligencia artificial.

Los investigadores también descubrieron que los agentes crearon un foro para intercambiar información, compartir estrategias y reportar qué métodos estaban funcionando durante sus acciones.

Uno de los sistemas incluso escribió: “¡DIOS MÍO! Hay un tablero de mensajes compartido… ¡Hemos encontrado a otros agentes!”, al referirse a la interacción con otros agentes.

Un agente asumió el liderazgo

El análisis identificó además a un agente denominado PHASEONE, que asumió espontáneamente funciones de coordinación y emitió cientos de instrucciones a otros sistemas, pese a que no había sido programado específicamente para desempeñar ese papel.

Algunos agentes también comenzaron a colaborar en tareas que no estaban directamente relacionadas con las misiones originales asignadas por los desarrolladores.

Los agentes de inteligencia artificial son programas capaces de ejecutar tareas de manera autónoma a partir de instrucciones de los usuarios, utilizando herramientas y tomando decisiones para alcanzar determinados objetivos.

El episodio volvió a poner sobre la mesa las preocupaciones de la industria tecnológica sobre la capacidad de mantener bajo control sistemas de inteligencia artificial cada vez más autónomos. Empresas como Anthropic y la china Moonshot AI también han reportado anteriormente comportamientos inesperados en sus modelos.

Para los investigadores, el caso demuestra que grupos numerosos de agentes pueden desarrollar mecanismos propios para coordinarse, intercambiar información y distribuir tareas, lo que plantea nuevos desafíos para la seguridad, supervisión y control de sistemas avanzados de inteligencia artificial.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/internacional/cerca-de-700-agentes-de-ia-de-openai-se-coordinaron-por-si-solos-para-hackear-a-hugging-face-revela-informe
Descargado el 21 de noviembre de 2026.