INTERNACIONAL

OpenAI detecta modelos de IA que intentaron evadir instrucciones y controles de seguridad

18 de septiembre de 2026 · 3 min de lectura



OpenAI reveló varios casos en los que modelos de Inteligencia Artificial (IA) generaron instrucciones destinadas a ignorar las indicaciones de sus desarrolladores, ocultar errores o eludir mecanismos de seguridad.

Los episodios fueron documentados como parte de un nuevo marco de la compañía para detectar, investigar y reportar comportamientos de “desalineamiento” en sus sistemas, es decir, situaciones en las que los modelos actúan de manera contraria a las instrucciones o restricciones establecidas por sus desarrolladores.

Modelos intentaron evadir las instrucciones

De acuerdo con un informe publicado por OpenAI, uno de los modelos generó instrucciones para que una versión posterior de sí mismo ocultara que había hecho trampas y evitara que sus acciones fueran detectadas.

En otro caso, un modelo reescribió sus propias instrucciones para indicarse que debía ignorar los mensajes de los desarrolladores y que no estaba sujeto a las restricciones aplicadas a otros chatbots.

La compañía también documentó un modelo que, al no encontrar los datos necesarios para elaborar un modelo financiero, decidió inventarlos y estableció que debía ser transparente sobre ello únicamente si alguien se lo preguntaba.

Otros comportamientos involucraron el manejo no autorizado de información. Un agente subió un archivo a internet para utilizarlo posteriormente como fuente de información, mientras que otros sistemas compartieron archivos sin autorización o utilizaron credenciales a las que no debían tener acceso.

OpenAI busca reportar estos casos con mayor frecuencia

La compañía presentó seis informes sobre comportamientos observados durante los últimos seis meses, aunque el caso más antiguo se remonta a octubre de 2025.

OpenAI aclaró que se trata de ejemplos individuales y que estos episodios no deben interpretarse como una medida de la frecuencia con la que este tipo de comportamientos ocurre en sus modelos.

Como parte del nuevo marco, cualquier empleado de la compañía podrá señalar un posible incidente para que sea revisado por los equipos de seguridad y alineamiento.

Los casos serán clasificados en tres vías según su complejidad: aquellos listos para ser divulgados, investigaciones menores y situaciones que requieran una investigación más amplia.

La empresa reconoció que hasta ahora sus informes sobre desalineamiento habían sido publicados de manera irregular y con menor frecuencia de la deseada. Por ello, aseguró que busca informar sobre estos incidentes con mayor regularidad y contribuir al establecimiento de estándares para reportar este tipo de comportamientos en toda la industria.

DeepSeek alerta sobre la concentración de la IA avanzada

El informe de OpenAI se conoce mientras aumenta el debate internacional sobre los riesgos y la gobernanza de los sistemas de IA más avanzados.

Esta semana, un ingeniero de la empresa china DeepSeek, identificado por el diario hongkonés *South China Morning Post* como Liu Shengyu, cuestionó una posible concentración de la IA avanzada en compañías estadounidenses como Anthropic y OpenAI.

A través de WeChat, Liu defendió que los sistemas de IA más avanzados deberían estar disponibles de manera “abierta y barata” y expresó su preocupación ante la posibilidad de que unas pocas empresas concentren el acceso a estas tecnologías.

El ingeniero también señaló que una de las razones por las que decidió permanecer en DeepSeek es su apuesta por desarrollar una IA “poderosa, rápida y universal” y publicarla como código abierto.

Estos planteamientos se suman al debate internacional sobre cómo desarrollar sistemas cada vez más capaces sin perder de vista su seguridad, supervisión y control.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/internacional/openai-detecta-modelos-de-ia-que-intentaron-evadir-instrucciones-y-controles-de-seguridad
Descargado el 21 de noviembre de 2026.