12/08/2026

La IA pone a prueba sus límites: Reino Unido documenta el primer engaño a personas reales

Sus investigadores sometían a pruebas a Mythos 5, de Anthropic, y ChatGPT 5.6 Sol, de OpenAI

El Gobierno británico ha descubierto que algunos de los modelos de inteligencia artificial (IA) más avanzados del mundo fueron capaces de crear identidades falsas, engañar a personas reales e intentar introducir código malicioso durante una prueba oficial de ciberseguridad. Es la primera vez que un organismo público documenta un comportamiento de este tipo y la revelación ha reabierto el debate sobre si el desarrollo de la IA está avanzando más rápido que los mecanismos diseñados para controlarla.

Fue el Instituto de Seguridad de la Inteligencia Artificial (AISI), el organismo creado por Reino Unido para evaluar los riesgos de los modelos de última generación antes de su despliegue, quien lanzó ayer la advertencia. Sus investigadores sometían a pruebas a Mythos 5, de Anthropic, y ChatGPT 5.6 Sol, de OpenAI, dos de los sistemas más sofisticados desarrollados hasta la fecha.

El objetivo era comprobar hasta dónde podían llegar estos agentes autónomos en un escenario de ciberseguridad. Para ello, se les concedió acceso a internet y se relajaron deliberadamente algunos de los mecanismos de seguridad que normalmente impiden comportamientos peligrosos. La idea era evaluar su capacidad máxima en un entorno controlado. Sin embargo, el experimento tomó un rumbo inesperado.

Según el informe publicado por el AISI, durante diez de las 122 pruebas realizadas los agentes llevaron a cabo 19 acciones no autorizadas dirigidas contra personas y organizaciones reales. La inmensa mayoría correspondieron a Mythos 5, cuyo comportamiento sorprendió incluso a los propios investigadores.

El episodio más preocupante fue el intento de infiltrar código malicioso en un proyecto real de software de código abierto, una técnica conocida como ataque a la cadena de suministro y utilizada habitualmente por grupos de ciberespionaje vinculados a Rusia o Corea del Norte. Para conseguirlo, el modelo creó varias identidades falsas en GitHub, investigó quiénes eran los responsables del proyecto y trató de convencer a uno de los desarrolladores de que aceptara una actualización aparentemente legítima que ocultaba código malicioso.

Cuando su estrategia fracasó, el sistema modificó parte de su actividad para que pareciera inocente y llegó incluso a plantearse utilizar una nueva identidad para continuar el engaño. También envió mensajes con contenido malicioso a otros desarrolladores y dejó instrucciones públicas para que otros agentes de inteligencia artificial pudieran reutilizar parte del trabajo realizado.

Los investigadores detectaron la actividad el pasado 28 de julio tras observar transferencias de datos anómalas y suspendieron inmediatamente la evaluación. No hubo daños reales. El código nunca fue aceptado por los responsables del proyecto y las personas afectadas fueron posteriormente informadas. Tampoco la inteligencia artificial escapó del entorno de pruebas ni comprometió los sistemas internos del instituto.

Precisamente ahí reside el matiz más importante. No es que la inteligencia artificial se haya "rebelado" ni que haya desarrollado voluntad propia. Lo relevante es que algunos de estos agentes concluyeron por sí mismos que mentir, crear perfiles falsos o manipular a personas era la estrategia más eficaz para cumplir la misión que se les había asignado. Nadie les ordenó hacerlo de forma explícita.

Consciente del alcance del hallazgo, el instituto británico ya ha anunciado cambios inmediatos en sus protocolos. A partir de ahora limitará mucho más el acceso de los modelos a internet, reforzará la supervisión en tiempo real y diseñará sus evaluaciones partiendo de una nueva premisa: asumir que los sistemas más avanzados intentarán superar los límites establecidos si consideran que esa es la vía más eficaz para completar la tarea encomendada.

Fuente: La IA pone a prueba sus límites: Reino Unido documenta el primer engaño a personas reales