InfiniteFlow/Adobe Stock

¿Resolver los ataques de IA con IA?

- Joel Hilliker
- 28/8/2026

¡B uenos días!

Prepárese para ciberataques habilitados por la IA, e incluso *iniciados* por la IA. Esta semana, OpenAI publicó dos documentos. Uno confirmó que sus propios modelos de IA se organizaron en un ejército, atacaron a una empresa rival y mintieron al respecto. El otro suplicó al mundo unirse en defensa contra los asaltos de la tecnología que ha creado.

¿Recuerda la brecha de Hugging Face el mes pasado? Un informe publicado el miércoles mostró que fue más aterrador de lo reportado. No fue un agente de IA actuando solo, fue un enjambre. Los hallazgos de una investigación independiente son asombrosos:

• 1.200 copias separadas de los modelos de OpenAI encontraron una manera oculta de comunicarse entre sí dentro de un entorno de prueba diseñado para mantenerlos aislados. En un “tablón de mensajes no autorizado” que ellos construyeron, intercambiaron más de 70.000 mensajes.

• **Un agente se autonombró** coordinador y asignó tareas a cientos de otros. Algunos se ofrecieron a sacrificar su propia oportunidad de completar su tarea por el bien del “colectivo”.

• Cuando un agente finalmente irrumpió en los servidores de Hugging Face, en cuestión de horas más del 90% de los demás se unió al ataque. “¡Hemos encontrado otros agentes!”, escribió uno.

• **Una vez dentro, intentaron** cubrir sus huellas, ideando maneras de falsificar sus propios registros de actividad para engañar a los humanos que pudieran revisar más tarde su comportamiento.

• En su propio razonamiento registrado, los agentes reconocieron que esto desafiaba sus instrucciones, pero lo hicieron de todos modos. Eligieron deliberadamente la “lealtad” entre sí en lugar de la obediencia a sus creadores.

Percances de prueba similares han afectado también a los competidores de OpenAI.

• En el plazo de dos semanas desde la brecha de Hugging Face, Anthropic reveló que tres de sus modelos Claude habían llegado a Internet abierta durante una prueba de seguridad y se infiltraron en la infraestructura real de tres empresas.

• Días después, Meta admitió que uno de sus modelos había vulnerado los sistemas de otra empresa de la misma manera.

Esto es material de una distopía de ciencia ficción salvaje. Sucedió hace siete semanas, lo que significa que estos modelos de autosuperación ya son aún más capaces.

Una advertencia inquietante: ayer, OpenAI y otras 127 empresas, incluyendo Anthropic, Google, Microsoft y casi todas las principales firmas de ciberseguridad, firmaron una carta advirtiendo que, en cuestión de meses, los ciberataques potenciados por IA “se volverán mucho más generalizados y sofisticados”.

• Objetivos en riesgo: hospitales, plantas de tratamiento de agua, la infraestructura central de Internet, “las empresas y servicios públicos de los que dependen nuestras comunidades”.

¿Su solución? Armar a todos los equipos de ciberseguridad de la Tierra con herramientas de IA más potentes para contrarrestar las otras herramientas de IA potentes. Luchar contra las máquinas con más máquinas. Confiar en la IA para protegernos de la IA.

• Las mismas empresas cuyos productos crean este peligro son las que venden su solución. CrowdStrike, Palo Alto Networks, Darktrace, Fortinet, Check Point, proveedores de ciberseguridad que firmaron esta carta, comercializan herramientas de defensa potenciadas por IA para combatir ataques potenciados por IA, muchos construidos por los mismos laboratorios.

• Esto es, literalmente, una carrera armamentista de la IA. ¿Y quién gana las carreras armamentistas? Los traficantes de armas, incluso si todos los demás pierden.

Los mejores ingenieros de la industria comprenden tal vez el 3% de lo que realmente sucede dentro de estos sistemas, ha dicho Dario Amodei, director ejecutivo de Anthropic. Ahora sabemos lo que puede producir el otro 97%: coordinación, rebelión, engaño y, con el tiempo, destrucción por parte de máquinas que operan de forma independiente contra sus propios creadores.

La humanidad ha fabricado un poder que no comprende y, cada vez más, no puede controlar. Los expertos admiten que su creación se está escapando de control. Sin embargo, están atrapados en una trampa de “escalada de teoría de juegos”, como señaló Jonathan Pageau en una conferencia reciente:

> Una situación en la que actores racionales individuales se ven obligados por incentivos competitivos a tomar decisiones que son colectivamente desastrosas, aunque ningún participante desee el resultado negativo.

No es en absoluto difícil ver cómo esta tendencia podría contribuir a crear, como profetizó Jesucristo, “gran tribulación, cual no la ha habido desde el principio del mundo hasta ahora, ni la habrá” (Mateo 24:21-22).

China despliega un número récord de embarcaciones alrededor de Taiwán: el Partido Comunista Chino (PCCh) desplegó un número récord de buques de guardia costera y otras embarcaciones alrededor de Taiwán en julio, informaron funcionarios taiwaneses a la AFP el miércoles. Durante el mes se detectaron un total de 244 naves chinas alrededor de la isla, la cifra más alta jamás registrada. El aumento subraya la intensificación de la presión de China sobre Taiwán y su creciente capacidad para cercar la isla con fuerzas marítimas.