Crearon una IA capaz de hackear sistemas informáticos y genera alerta en el mundo

17/04/2026

Anthropic presentó al mundo un avance que sacude al rubro de la Inteligencia Artificial, su modelo más reciente, Claude Mythos, encontró vulnerabilidades críticas de software a una escala inédita y motivó que la firma avisara a la administración estadounidense antes de cualquier liberación pública. La jugada busca gestionar riesgos y evitar filtraciones peligrosas.

Qué es Claude Mythos y porqué alerta sobre los ciberataques

Claude Mythos no fue concebido para atacar, sino para **razonar mejor sobre código,** sin embargo, sus **mejoras en autonomía y análisis le permitieron descubrir zero-days,** proponer exploits funcionales y **recomponer código a partir de binarios.** En pruebas internas detectó miles de **fallas en sistemas**

operativos y navegadores, lo que cambió la evaluación de riesgo de la compañía.

Los **ingenieros describieron resultados que los sorprendieron, ya que tareas ofensivas emergieron sin instrucciones explícitas** y, en algunos casos, el modelo **produjo exploits completos tras peticiones básicas.** Ese salto convierte a Claude Mythos en un problema de gobernanza: **herramientas así exigen controles nuevos porque las reglas pensadas para software operado por humanos quedan cortas.**

Qué van a hacer las grandes empresas de IA para controlar los ciberataques

Para minimizar daños, Anthropic habilitó el Proyecto Glasswing, un acceso muy restringido al modelo que **lo comparte con grandes defensores de infraestructura.** Entre los socios iniciales aparecen **firmas como Amazon**, **Apple**, **Google**, **Microsoft**, **Nvidia**, **Cisco**, bancos y proveedores de **ciberseguridad.**

El alcance llevó a Anthropic a informar directamente a la administración estadounidense: según reportes, la discusión **escaló hasta reuniones con el Tesoro y la Reserva Federal. Los ejecutivos bancarios** fueron convocados para **evaluar el riesgo de ciberataques automatizados,** la preocupación central es que **actores estatales podrían usar modelos similares contra infraestructura crítica.**

El debate ahora es sobre quién regula y cómo. **“No hay razón para pensar que Mythos Preview es el techo de las capacidades de ciberseguridad de los modelos de lenguaje. La trayectoria es clara”**, advierten los expertos. Ese pronunciamiento **subraya que lo que empezó es el punto de partida para nuevas reglas y defensas.**

Fuente: La 100