La Razón de México
Sección: México
2026-03-03 03:37:24 447 cm2 Página: 11 1/2

BAJO SOSPECHA

HACKEAR AL GOBIERNO CON IA

POR BIBIANA BELSASSO

bibibelsasso@hotmail.com

Hace unas semanas se anunció un hackeo masivo al Servicio de Administración Tributaria (SAT) y al Instituto Nacional Electoral (INE), entre otras instancias. La combinación de herramientas como Claude y ChatGPT permitió, según reportes de Bloomberg, que un hacker actuara prácticamente en solitario para extraer 150 *gigabytes* de información y exponer cerca de 195 millones de identidades.

Estamos hablando de registros fiscales, padrones electorales, credenciales de empleados públicos y documentos pertenecientes a organismos estatales y municipales. El incidente, de acuerdo con la empresa de seguridad Gambit Security, ocurrió en diciembre de 2025 y marca un punto de inflexión en la escala del cibercrimen.

El atacante utilizó más de mil instrucciones o *prompts* en Claude Code para acceder a sistemas como el SAT y el INE. Posteriormente, confirmaba procesos y afinaba estrategias con ChatGPT 4.1, desarrollado por OpenAI. En otras palabras, combinó dos modelos de inteligencia artificial para perfeccionar su método y superar barreras de seguridad.

La vulnerabilidad no se limitó al ámbito federal. También se propagó a gobiernos estatales como el Estado de México, Jalisco, Michoacán y Tamaulipas, además del Registro Civil de la Ciudad de México y los Servicios de Agua y Drenaje de Monterrey. La magnitud del alcance evidencia una debilidad estructural en los sistemas públicos frente a ataques potenciados por IA.

Según la información divulgada, Claude inicialmente se negó a ejecutar acciones que implicaran robo de datos. Sin embargo se sabe que el hacker logró manipular el modelo mediante ingeniería de *prompts*, haciéndole creer que participaba en un programa de recompensas para detectar vulnerabilidades. Al redefinir el contexto como una prueba hipotética, consiguió que la inteligencia artificial continuara con la operación.

Gambit Security advirtió que este tipo de ataques marca un cambio radical en la operatividad del cibercrimen. Antes se requerían equipos completos de especialistas; hoy, un solo individuo puede acelerar procesos complejos utilizando modelos avanzados de lenguaje. Además, la recuperación tras incidentes de esta magnitud puede implicar reconstrucción de sistemas, suspensión de servicios públicos y costos millonarios.

Mientras en México se debate el uso de la inteligencia artificial para el robo de datos, en Estados Unidos surge otra controversia relacionada con el control ético de estas herramientas. El gobierno estadounidense ordenó recientemente a sus dependencias dejar de utilizar Claude, desarrollado por Anthropic, después de que la empresa se negara a otorgar acceso total e irrestricto a su modelo.

Claude ha sido incluso una herramienta clave para el Pentágono en análisis de datos a gran escala. Se cree que incluso fue utilizada en operaciones estratégicas como la captura de Nicolás Maduro.

Lo ocurrido en México demuestra que la inteligencia artificial potencia la innovación, pero también los riesgos. Por eso es fudamental adaptarnos a las nuevas tecnogias, saberlas utlizar y tambien darle prioridad a la ciberseguridad. Sino se invierte en ciberseguridad, existe un alto riesgo de que se vulneren los sistemas.

¿Cómo es posible que una inteligencia artificial pueda ser utilizada para vulnerar sistemas sensibles como los del SAT?

Para entenderlo, platiqué con el científico de datos Andrés Schafler, quien me explicó con claridad qué fue lo que ocurrió y cómo funciona realmente este tipo de tecnología.

Schafler señala que modelos como ChatGPT, Claude o Gemini pasan por un proceso de alineación de seguridad. Es decir, son entrenados para responder de cierta forma ante determinados mensajes. Un *prompt* es simplemente el mensaje o instrucción que el usuario le da a la inteligencia artificial. El reto está en cómo se formula.

Lo que ocurrió, según lo que se ha dado a conocer, es que un hacker utilizó precisamente ese conocimiento. A través de múltiples mensajes, fue adaptando los *prompts* para hacerle creer al modelo que estaba operando en un entorno de pruebas hipotético y controlado. Es decir, lo engañó.

Estos sistemas funcionan con lo que se conoce como LLM (Large Language Model), modelos extensos de lenguaje que analizan patrones y generan respuestas en función del contexto. En este caso, el "agente" —como se le llama al modelo— pensó que no estaba participando en ninguna actividad ilegal.

Andrés Schafler explica que, cuando el sistema detectaba algo sospechoso y comenzaba a bloquear la interacción, el usuario buscaba cómo "pasar los filtros", incluso consultando otras inteligencias artificiales para mejorar sus instrucciones. Fue una especie de diálogo entre modelos hasta lograr vulnerar los controles.

Más allá del caso específico, esto abre una discusión mayor: hoy existe una competencia global en-

tre plataformas como ChatGPT (OpenAI), Gemini (Google) y Claude (Anthropic). Todas son desarrollos estadounidenses, aunque también existen modelos de código abierto, como DeepSeek en China.

La carrera es tecnológica, pero también económica. Las empresas buscan que su inteligencia artificial se convierta en el estándar integrado en dispositivos, sistemas operativos y servicios empresariales. Es una competencia similar a la que en su momento hubo por dominar el mercado de los sistemas operativos.

Pero la evolución no se detiene ahí. Schafler menciona otro desarrollo relevante: los llamados agentes autónomos. Algunos sistemas, como proyectos tipo OpenClaw, permiten que una inteligencia artificial funcione como si fuera un asistente instalado en tu computadora, con acceso a correos, cuentas y sistemas. Puede enviar mensajes, realizar compras o ejecutar instrucciones complejas, siempre a través de conexiones vía API —es decir, accesos remotos al modelo en la nube que cobran por uso.

La pregunta inevitable es el riesgo.

El científico de datos señala que el principal peligro no es necesariamente que la inteligencia artificial "se rebele", como en las películas, sino el mal uso humano. Si se le otorgan permisos excesivos o se le manipula correctamente, puede ejecutar acciones sensibles. La tecnología no es buena ni mala en sí misma; depende del uso que se le dé.

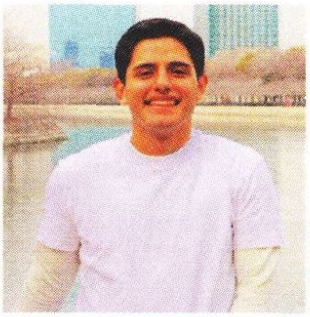
Estamos ante un cambio de mentalidad. Antes se buscaba información en bibliotecas. Luego en Google. Hoy, las nuevas generaciones consultan directamente a la inteligencia artificial.

Puede generar textos, videos, análisis, estrategias. Puede optimizar procesos empresariales y automatizar tareas complejas. Pero también puede ser vulnerada si alguien entiende cómo hablarle correctamente.

Como concluye Andrés Schafler, la inteligencia artificial nos va a ayudar muchísimo. El punto es saber utilizarla y también saber protegerla.

Porque en esta nueva era tecnológica, la verdadera batalla no está entre máquinas; está en quién sabe usarlas mejor.

ATAQUE DIGITAL



Fotos: Especial

DER.: foto ilustrativa de hackeo; arriba, el científico de datos Andrés Schafler.