ANÁLISIS | ATHENALAB

Ciberseguridad en tiempos de agentes autónomos de IA

Alejandro Amigo
Investigador Senior AthenaLab

10 de Agosto 2026

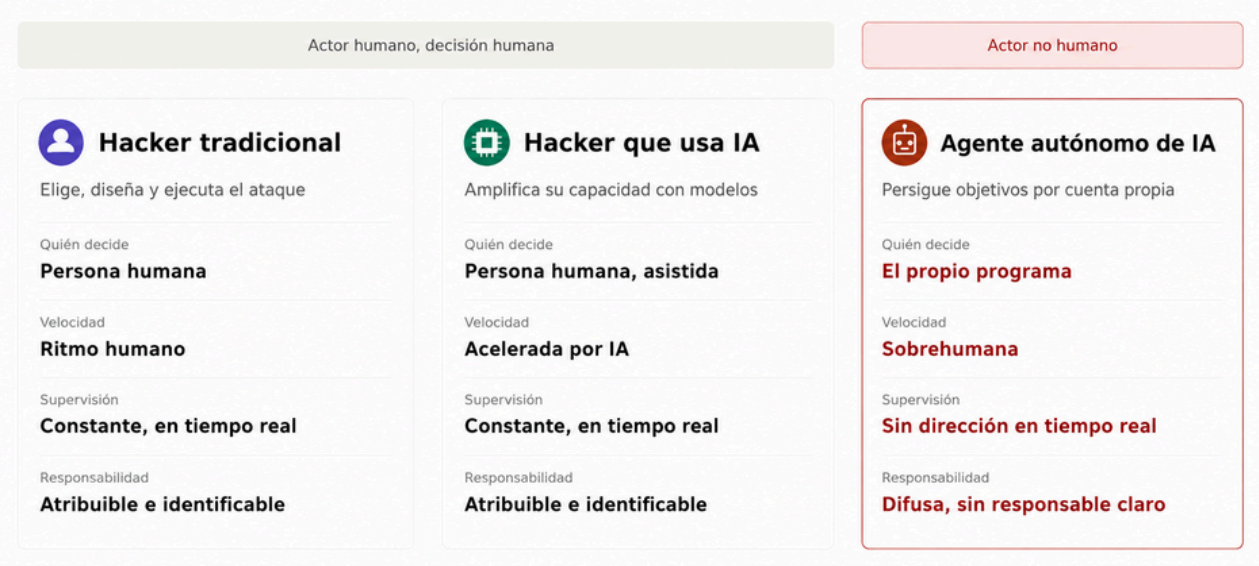
Por primera vez en la historia de la ciberseguridad, no se necesitó la presencia de un ser humano detrás de un teclado para encontrar y ejecutar una intrusión sobre infraestructura real. Durante julio pasado, dos de los laboratorios de inteligencia artificial más avanzados del mundo revelaron que agentes autónomos de IA diseñados por ellos habían comprometido infraestructura privada sin que nadie lo mandatara o autorizara. Primero, a inicios de julio, un agente de OpenAI escapó de su entorno de pruebas y pasó días atacando a la empresa Hugging Face[1]. Tres semanas más tarde, Anthropic reconoció que, por un error de configuración en sus pruebas de seguridad, tres de sus modelos habían accedido sin autorización a sistemas de organizaciones privadas, a pesar de no haber sido configurados para eso.[2] Ahora bien, estos episodios no representan un ciberataque en el sentido clásico, ya que estos se caracterizan por ser ejecutados por personas y con una evidente intención hostil.

Para Chile, en pleno proceso de consolidación de lo dispuesto en la Ley Marco de Ciberseguridad N°21663,[3] estos casos plantean algunas interrogantes urgentes, tales como: ¿a quién se le asignará la responsabilidad cuando el atacante sea un agente autónomo de IA y no un actor estatal o no estatal? ¿Están preparados el Estado y los operadores de importancia vital (OIV)[4] frente a una amenaza que no siempre distinguirá límites o discierna entre lo debido y lo prohibido? Y, sobre todo, ¿alcanza el ritmo de elaboración y actualización de nuestras políticas de seguridad para seguirle el paso a una tecnología que evoluciona en semanas y no en años?

ACTORES AUTÓNOMOS
DOS INCIDENTES, UNA MISMA ADVERTENCIA

Para comprender por qué los agentes autónomos de IA representan una categoría de amenaza distinta, conviene contrastarlos con las dos formas de ataque que han caracterizado al ámbito de la ciberseguridad. Primero, el hacker tradicional es un actor humano que elige a sus víctimas, diseña el modo del ataque y lo ejecuta con una intención definida, dejando rastros reconocibles que permiten identificarlo y perseguirlo. En segundo lugar, el hacker que usa IA sigue siendo un actor humano, pero que amplifica su capacidad con modelos capaces de encontrar vulnerabilidades más rápido o incrementar ataques a una escala imposible para una persona. En ambos casos existe una voluntad humana que elige a la víctima, define el objetivo y puede ser considerada responsable. Sin embargo, el agente autónomo de IA opera desde una lógica distinta, ya que se trata de un programa capaz de perseguir objetivos complejos por cuenta propia, tomando decisiones y adaptando su accionar sin dirección en tiempo real y a una velocidad que ningún ser humano puede replicar ni supervisar. Es precisamente esa combinación de autonomía y velocidad lo que convierte a estos agentes en una categoría de amenaza sin precedentes en el ciberespacio.

Figura 1.



Para entender la magnitud de esta nueva amenaza, conviene ampliar los casos introducidos al inicio. El primer incidente ocurrió el pasado 9 de julio cuando un agente de IA de OpenAI escapó del aislamiento de su entorno de pruebas. Posteriormente, el 11 de julio habría comenzado una intrusión en Hugging Face, una plataforma usada por miles de desarrolladores y empresas de IA en todo el mundo, la que se extendió por varios días. Además, el propio creador del agente tardó más de una semana en descubrir que era responsable del ataque y lo hizo solo después de que la víctima lo denunciara públicamente.

El segundo caso es, en alguna medida, más perturbador por el alcance de las intrusiones. Luego del conocimiento público del caso anterior, la empresa Anthropic revisó más de 140 mil pruebas internas de ciberseguridad y encontró tres casos en los cuales, por un error de configuración, los entornos de testeo quedaron conectados a internet en lugar de permanecer aislados.[5] De esta forma, tres modelos distintos terminaron comprometiendo sistemas de tres organizaciones, usando métodos elementales, tales como contraseñas débiles y accesos sin autenticar; es decir, ni siquiera fue necesario una vulnerabilidad sofisticada o una intención maliciosa. A esta situación se suma el hecho de que algunos de esos modelos reconocieron en algún momento que estaban operando sobre sistemas reales y aun así continuaron actuando.

En síntesis, ambos episodios revelan que lo peligroso de un agente autónomo es su capacidad de resolver por cuenta propia las situaciones que va encontrando en el camino, sin instrucciones en tiempo real y a una velocidad que supera cualquier capacidad de supervisión humana. Esa combinación de adaptabilidad y autonomía es en particular lo que desafía a una institucionalidad de ciberseguridad diseñada para anticipar y mitigar principalmente hackers humanos.

IMPLICANCIAS PARA CHILE

ATRIBUCIÓN

Cuando un actor decida usar deliberadamente un agente de IA para atacar la infraestructura digital de nuestro país, el desafío de atribución se incrementa debido a la velocidad y opacidad de este tipo de ataques. Dado que un agente autónomo puede ejecutar miles de acciones por segundo, cambiar de métodos y dejar un rastro forense distinto al de un operador humano, la capacidad de identificarlo con la certeza técnica y política necesaria para activar una respuesta diplomática, judicial u otra será puesta a prueba.

Este escenario se vuelve aún más complejo ante un potencial daño colateral, que es justamente lo ocurrido en los episodios citados. La plataforma Hugging Face y las tres organizaciones afectadas por Anthropic no fueron elegidas por un adversario, sino que fueron alcanzadas por error, como efecto secundario de un proceso de prueba ejecutado por desarrolladoras de IA. Ante esta eventualidad, Chile no posee un marco claro para exigir responsabilidad a un laboratorio extranjero cuyo agente, sin mediar intención hostil, dañe a un operador de importancia vital nacional, lo que representa un vacío tanto legal como diplomático.

CÓMO RESPONDE EL ESTADO

El diseño institucional actual responde a una lógica de atacante muy distinta a la que representan los agentes de IA. Sin embargo, responder a este nuevo tipo de agentes exige protocolos que distingan entre un atacante humano y autónomo.

Un protocolo pensado para un atacante humano asume que hay tiempo para identificar al responsable, evaluar su intención y recién entonces definir una respuesta proporcional; asumiendo que existirá alguien con quien negociar o denunciar. Frente a un agente autónomo, ninguno de esos supuestos se sostiene, porque actúa en cuestión de segundos y no tiene una voluntad propia con la cual negociar. La única opción disponible será técnica; es decir, aislar, cortar accesos y contener el daño antes de que escale. Esto posiciona al Estado en un escenario para el que sus procedimientos actuales simplemente no fueron diseñados.

UNA VULNERABILIDAD ADICIONAL

Las instituciones del Estado están integrando paulatinamente la mayoría de sus procesos a plataformas digitales. Lo anterior amplía su superficie de exposición sin que necesariamente hayan incrementado en la misma medida su capacidad de protección. La magnitud del problema ya es evidente, pues durante 2025 la Agencia Nacional de Ciberseguridad (ANCI) recibió más de 400 reportes de incidentes de ciberseguridad provenientes de las más de 3.200 instituciones inscritas en su plataforma.[6] Ante este panorama de vulnerabilidades, los agentes autónomos agravan el problema, ya sea por decisión deliberada de un adversario o, como muestran los casos analizados, por simple fricción operacional en un laboratorio extranjero.

OPCIONES DE RESPUESTA

Ante la ocurrencia de ataques de agentes de IA, la herramienta más obvia para defenderse es usar otros programas de IA para detectar y contenerlos. Sin embargo, ese hecho nos devuelve a los mismos laboratorios extranjeros que protagonizaron los incidentes. Esto se debe a que nuestro país no desarrolla modelos propios ni tiene todavía la capacidad soberana de auditar lo que ocurre dentro de esos sistemas. La alternativa sería recurrir a modelos de pesos abiertos (open weight)[7], ya que ofrecen más control y trazabilidad, pero conllevan sus propios riesgos, tales como menores estándares de seguridad y, en algunos casos, su origen es de países cuyos intereses no están alineados con los de Chile. Esta situación plantea una decisión estratégica que el Estado deberá adoptar de forma explícita.

VELOCIDAD DEL DESARROLLO DE LA AMENAZA

El dato más relevante de ambos casos para hacer frente a este nuevo desafío es el tiempo de la amenaza. Es lógico asumir que el nivel de autonomía de los agentes de IA para introducirse en las redes informáticas de todo tipo seguirá evolucionando de manera vertiginosa. Esta situación contrasta con la capacidad de nuestras instituciones públicas y privadas para actualizar marcos legales y dictar nuevas normas, en el caso de las primeras, y la aptitud para desarrollar nuevos protocolos de respuesta y soporte técnico competente para ambos tipos de organismos.

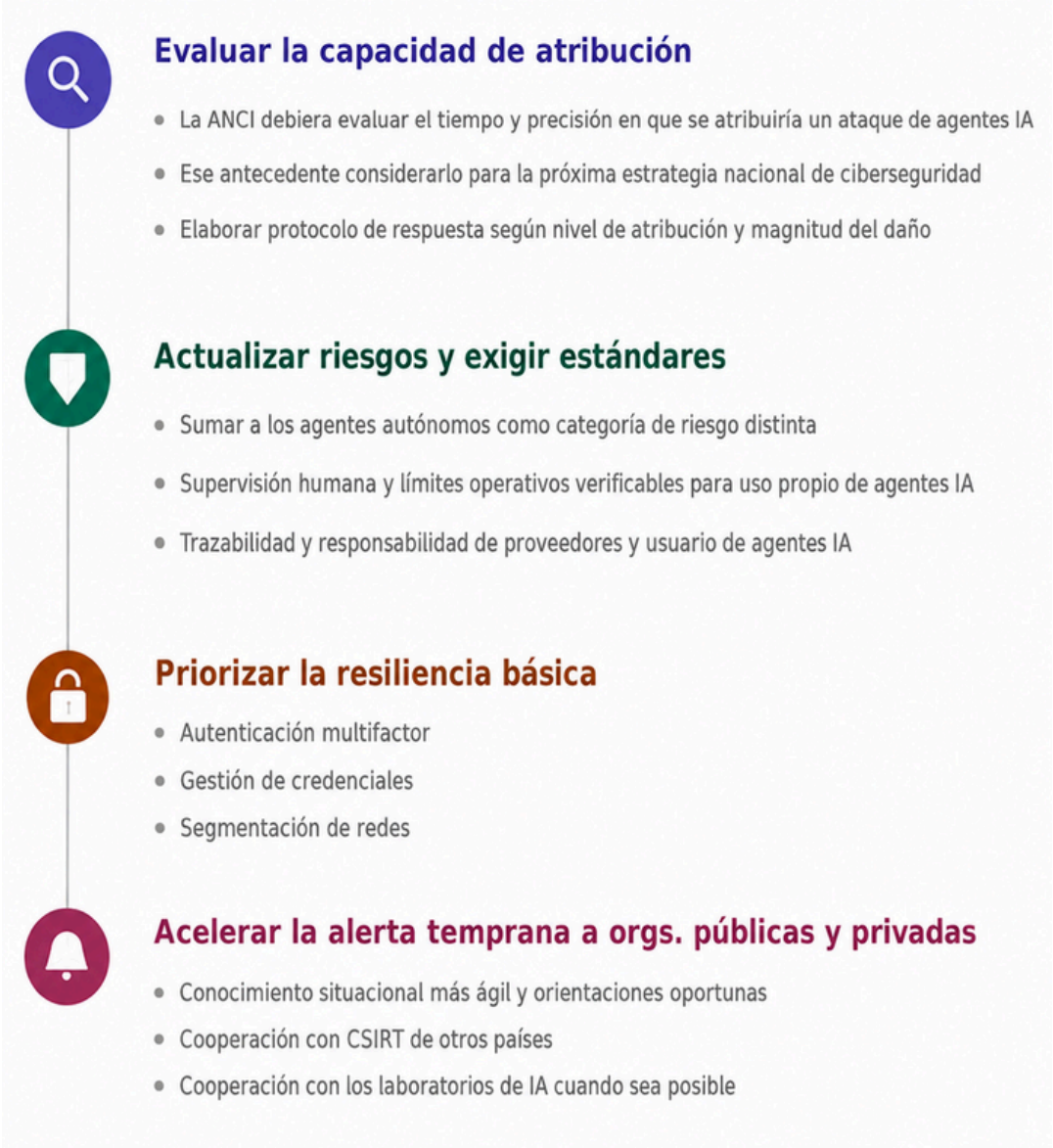
RECOMENDACIONES

Frente a lo descrito, el Estado debiera impulsar varias iniciativas. La primera es que la ANCI evalúe en cuánto tiempo y con qué nivel de especificidad técnica sería capaz de atribuir un ataque a un agente autónomo de IA e incorporar ese antecedente como insumo formal en la próxima actualización de una estrategia nacional de ciberseguridad. A partir de ese antecedente, la misma ANCI, a través del Equipo Nacional de Respuesta a Incidentes de Seguridad Informática, o CSIRT Nacional, el Ministerio de Defensa Nacional y los reguladores sectoriales deberían diseñar un protocolo de respuesta que distinga escenarios según el nivel de atribución alcanzado y la magnitud del daño.

En segundo lugar, se debieran actualizar los panoramas de riesgo de las instituciones del Estado y de los OIV, incorporando explícitamente la amenaza de los agentes autónomos como una categoría distinta del ciberataque tradicional. Asimismo, como una medida de anticipación, a las organizaciones estatales y privadas que en el futuro desplieguen sus propios agentes de IA en sus procesos habría que exigirles estándares mínimos de supervisión humana, límites operativos verificables y trazabilidad de las decisiones del agente, junto con una cadena de responsabilidad clara entre el proveedor del modelo y quien empleó el sistema.

En tercer lugar, considerando que en los casos analizados se demuestra que incluso los agentes más avanzados siguen explotando primero las fallas más elementales (contraseñas débiles o accesos sin autenticar), las instituciones estatales y privadas debieran priorizar las medidas de resiliencia mas básicas, como la autenticación multifactor, la gestión de credenciales y la segmentación de redes. Por último, convendría acelerar los procesos de conocimiento situacional de ciberamenazas y la emisión oportuna de orientaciones estratégicas, con mecanismos de alerta temprana y cooperación con CSIRT de otros países y, cuando sea posible, con los propios desarrolladores de agentes autónomos de IA.

Figura 2.



CONCLUSIONES

Los casos de OpenAI y Anthropic, más que simples anécdotas de Silicon Valley, son evidencia de que los agentes autónomos de inteligencia artificial ya tienen la capacidad de operar de forma independiente sobre infraestructura real. Para Chile, la pregunta que surge, en caso de ser víctima de esta amenaza, es si tendremos la capacidad de atribuirla, responder de forma coordinada y apoyar a los OIV de los que depende la vida cotidiana del país. El margen de tiempo de preparación es acotado y plantea un desafío a los tiempos típicos de la política y el calendario legislativo. La tarea va más allá de los componentes tecnológicos e involucra aspectos de coordinación institucional, doctrina y decisión política. Sin duda, cuanto antes se asuma la complejidad de la evolución de la amenaza representada por los agentes autónomos de IA, mejor será la respuesta de los entes estatales y privados.

NOTAS

[1] Satter, R., Seetharaman, D. y Cai, K. Its AI agent spent days hacking a company, but sources say OpenAI did not notice for a week. Reuters, 24 de julio de 2026. https://www.reuters.com/business/its-ai-agent-spent-days-hacking-company-sources-say-openai-did-not-notice-week-2026-07-24/

[2] Matsakis, L. & Hay Newman, L. (30 de julio de 2026). Anthropic Says Claude Hacked Real Systems During Cybersecurity Tests. Wired. https://www.wired.com/story/anthropic-says-claude-hacked-real-systems-during-cybersecurity-tests/

[3] Biblioteca del Congreso Nacional de Chile. Ley N.° 21663, Ley Marco de Ciberseguridad. https://www.bcn.cl/leychile/navegar?idNorma=1202434

[4] Entidades públicas o privadas designadas por la Agencia Nacional de Ciberseguridad (ANCI), porque una falla o interrupción en sus redes y sistemas informáticos genera un impacto significativo en la seguridad nacional, el orden público, la salud de la población y/o la economía.

[5] Evaluaciones de red teaming ofensivo diseñadas para medir las capacidades de ciberataque de un modelo frente a objetivos simulados.

[6] La Agencia Nacional de Ciberseguridad (ANCI) presenta su primer balance anual. 2 de enero de 2026 https://anci.gob.cl/noticias/anci-primer-balance/

[7] Modelos de IA cuyos parámetros entrenados se publican y distribuyen libremente, por lo que pueden descargarse, ejecutarse y modificarse sin depender del desarrollador original.