Anthropic, la empresa de IA que se enfrentó al Pentágono en EE.UU. y por qué esto nos concierne a todos

Tiempo de lectura: 12 min.

Dalia Ventura

Mientras el mundo observaba la operación estadounidense en Venezuela y cómo la guerra con Irán se hacía inevitable, en Washington se iba fraguando una batalla que advertía que el futuro profetizado durante siglos ya es presente.

Una empresa de Inteligencia Artificial (IA) de Silicon Valley le dijo no al Pentágono y este la trató como si fuera enemiga del Estado. No obstante, su tecnología de IA seguía siendo usada, porque el ejército de EE.UU. no podía darse el lujo de prescindir de ella.

Eso es lo que pasó entre Anthropic y el Departamento de Defensa en las últimas semanas. Y aunque suena a disputa corporativa, es mucho más que eso.

Es la primera vez que una empresa de IA enfrenta a un aparato militar, negándose a eliminar los límites éticos de su tecnología.

El enfrentamiento dejó preguntas en el aire que nos atañen a todos: ¿Hasta dónde ya estamos delegando decisiones irreversibles y letales en las máquinas? ¿Quién decide cómo se usa la IA?

No son preguntas retóricas. Expertos de de la Universidad de Oxford advierten que este episodio "revela brechas de gobernanza de larga data en la integración de la IA en operaciones militares, brechas que preceden a esta administración y sobrevivirán a la controversia actual".

¿Por qué, si la humanidad lleva tanto tiempo temiendo llegar a este punto, aún hay tal vacío en la gobernanza de la IA?

Se trata de un vacío que Logan Graham, líder del equipo rojo de Anthropic, el cual analiza los peores escenarios de la tecnología, desde ciberataques hasta amenazas de bioseguridad, conoce de cerca.

"La intuición de alguna gente, por haber crecido en un mundo pacífico, es que en algún lugar hay una sala llena de adultos que saben cómo arreglar todo", le dijo a la revista Time.

"No existen esos grupos de adultos. Ni siquiera existe la sala. Tú eres responsable".

Retomemos lo que ocurrió.

Una llamada incómoda

En algún momento de la operación que culminó el pasado 3 de enero con la captura del entonces presidente de Venezuela, Nicolás Maduro, una herramienta de IA llamada Claude estuvo presente, procesando datos y ayudando a tomar decisiones.

Así lo reportaron de manera independiente el diario Wall Street Journal y el sitio web Axios, citando fuentes con conocimiento directo de los hechos, y lo reafirmó posteriormente la revista Time, que publicó un extenso perfil de Anthropic, la empresa de San Francisco que creó Claude.

Ni el Departamento de Defensa ni Anthropic lo confirmaron oficialmente. Pero lo que sucedió después está documentado, y dice más que el hecho mismo.

Tras la captura de Maduro, un ejecutivo de Anthropic contactó a Palantir -la empresa de análisis de datos que actúa como intermediaria tecnológica entre Silicon Valley y el gobierno estadounidense- y preguntó: ¿fue usado nuestro software en esa operación?

La pregunta encendió alarmas en Washington.

Emil Michael, el subsecretario de Defensa y jefe de tecnología del Pentágono, explicó que les generó una preocupación profunda: ¿podría Anthropic, en un conflicto futuro, "apagar su modelo en medio de una operación" -activar algún

mecanismo de rechazo- "y poner vidas en riesgo"?

Anthropic disputa esa lectura: la empresa dice que jamás intentó limitar el uso del Pentágono en un caso concreto y que la pregunta fue rutinaria.

Lo que siguió fue una escalada a cámara rápida.

El Pentágono exigió que Anthropic entregara acceso irrestricto a su tecnología para "todos los usos legales". Anthropic se negó.

Pete Hegseth, el secretario de Defensa de Trump, designó a Anthropic como un "riesgo en la cadena de suministro", una etiqueta que históricamente se reserva para empresas vinculadas a adversarios extranjeros como Huawei o Kaspersky, no para compañías estadounidenses que simplemente discrepan con el gobierno.

Anthropic demandó al Pentágono por exceder su autoridad y sus salvaguardas éticas, violando derechos básicos. Varios expertos legales consideran que la empresa tiene opciones sólidas de ganar en los tribunales.

El presidente Donald Trump, por su lado, ordenó a todas las agencias federales que dejen de usar la tecnología de Anthropic.

Y remató la polémica con un mensaje en la plataforma Truth Social, escrito todo en mayúsculas: "Estados Unidos nunca permitirá que una empresa de izquierda radical y woke dicte cómo combate y gana guerras nuestro gran ejército".

En su vocabulario y el de sus seguidores, 'woke' es el insulto máximo, una etiqueta despectiva para describir ideas o políticas progresistas sobre identidad, desigualdad o justicia social.

Quizás el calificativo es adecuado: Anthropic, efectivamente, se empeña en ser una empresa 'virtuosa'.

Líneas rojas

Anthropic tiene una historia sui géneris.

Fue fundada en 2021 por exinvestigadores de OpenAI con la premisa explícita de que la inteligencia artificial representa uno de los mayores riesgos existenciales para la humanidad y que, precisamente por eso, es mejor que quienes la desarrollen sean personas comprometidas con hacerlo de manera segura.

En julio de 2025, firmó un contrato de US$200 millones con el Departamento de Defensa, el primero de su clase: un laboratorio de IA que integra sus modelos en flujos de trabajo de misiones en redes clasificadas.

El CEO de Anthropic, Dario Amodei, lo justificó en un ensayo publicado en enero de este año.

Anthropic, escribió, apoyaba a las fuerzas militares y de inteligencia estadounidenses porque "la única manera de responder a las amenazas autocráticas es igualarlas y superarlas militarmente".

Y añadió: "La formulación a la que he llegado es que debemos usar la IA para la defensa nacional en todas las formas, excepto en aquellas que nos harían más parecidos a nuestros adversarios autocráticos".

En concordancia, el contrato con el Pentágono trazaba dos "líneas rojas": Claude no podría usarse para vigilancia masiva doméstica ni para armas completamente autónomas.

Esos límites infranqueables no son arbitrarios; se sustentan en un documento de la empresa que funciona como su "alma".

Su objetivo declarado es "evitar catástrofes a gran escala", incluyendo la posibilidad de que la IA sea usada por un grupo humano para "tomar el poder de manera ilegítima y no colaborativa".

Amodei también argumentó ante el Pentágono que "los sistemas de IA de vanguardia simplemente no son lo suficientemente confiables como para impulsar armas totalmente autónomas".

En sentido estricto, no hablamos de armas que decidan por sí solas a quién matar; en este contexto, "autonomía" significa que un sistema pueda cumplir ciertos objetivos por su cuenta -o con mínima supervisión humana- en entornos complejos.

Pero sí existen sistemas automatizados que ayudan a tomar decisiones sobre ataques. Y los expertos en inteligencia artificial advierten de un problema conocido como "sesgo de automatización": cuando las reglas de uso son vagas, los humanos tendemos a confiar en las recomendaciones de la máquina más de lo que deberíamos.

La IA no reemplaza el juicio humano de golpe: lo va erosionando poco a poco, hasta que el operador deja de cuestionarla.

En una situación tensa, si el sistema -que sabes que analizó una cantidad inmensa de información- señala en la pantalla unos pocos píxeles como un objetivo urgente, es fácil aceptar su recomendación sin dudar lo suficiente.

O si un sistema de reconocimiento facial señala a alguien en medio de la multitud, es probable que un agente de seguridad confíe en el resultado y proceda al arresto.

Hay precedentes concretos: en múltiples ocasiones documentadas, varios departamentos de policía en EE.UU. terminaron arrestando personas equivocadas.

Eso resuena con la otra línea roja que enfureció al gobierno de Trump, la de la vigilancia masiva, que toca la vida cotidiana de personas que no están en ninguna zona de guerra.

Cabe anotar que Anthropic se opuso específicamente a la vigilancia masiva de ciudadanos estadounidenses. Su postura no es universalista. Pero el principio que la sustenta sí tiene un alcance más amplio.

Y cobra urgencia porque, en paralelo a este conflicto, el gobierno estadounidense anunció planes de usar IA a través de Palantir para apoyar las operaciones de ICE -la agencia de inmigración- rastreando ubicaciones en tiempo real e historial financiero de personas indocumentadas.

La vigilancia masiva, en distintos grados y sobre distintas poblaciones, ya existe. La pregunta ya no es si ocurre, sino cuántos controles quedan sobre cómo se usa.

En ese contexto, las "líneas rojas" de Anthropic no son solo filosofía corporativa: son, por ahora, uno de los pocos frenos concretos que existen.

El problema es que las restricciones son tan sólidas como el mecanismo que las hace cumplir.

Y cuando el Pentágono rechazó esos límites y exigió acceso irrestricto, Anthropic se encontró sola sosteniendo su postura, sin un marco legal que la respaldara, sin regulación internacional que la protegiera, con solo sus cláusulas contractuales

como escudo.

¿Qué es "legal"?

La reticencia del Departamento de Defensa a que una empresa privada le imponga límites es, para muchos, justificada.

Aunque las operaciones militares recientes en ciudades estadounidenses, Venezuela e Irán se llevaron a cabo con una mínima consulta al Congreso, el uso de la IA es tan crítico que debe estar regulado por leyes aprobadas por representantes elegidos democráticamente, opinan algunos.

Desafortunadamente, el poder legislativo no ha legislado al respecto.

Así, el hecho de que el Pentágono exija la libertad de usar a Claude para "todos los usos legales" suena razonable hasta que se pregunta qué es, exactamente, legal en este ámbito.

No existe una definición consensuada en el derecho internacional sobre qué constituye un arma letal autónoma.

El derecho internacional humanitario -las reglas que rigen los conflictos armados desde los Convenios de Ginebra- fue construido en torno a decisiones humanas: un soldado que aprieta un gatillo, un comandante que da una orden.

Esos marcos no contemplan sistemas que detectan, seleccionan y eliminan objetivos con mínima intervención humana directa o sin ella.

Es lo que los expertos llaman "vacío de responsabilidad": una deficiencia crítica en la que los marcos legales existentes no logran determinar quién responde cuando un sistema autónomo comete una infracción.

Si un dron con IA mata a civiles, ¿quién responde? ¿El programador? ¿El comandante? ¿La empresa que fabricó el sistema?

El derecho internacional no tiene una respuesta clara. Y en ausencia de esa respuesta, "uso legal" significa, en la práctica, lo que cada Estado decida que significa.

En este contexto, surge una pregunta delicada: ¿llega esta discusión a tiempo? La respuesta quizás es: a tiempo para ser preventiva, no; a tiempo para ser útil,

todavía sí.

El debate formal sobre las armas autónomas comenzó en 2013. Once años después, el resultado son guías voluntarias.

En 2024, durante una conferencia internacional en Viena, el ministro de Exteriores de Austria urgió a avanzar con una frase inquietante: "Este es el momento Oppenheimer de nuestra generación".

Aludía al momento en el que la humanidad tomó conciencia del poder destructivo de la bomba atómica: como entonces, la tecnología ya existe y ahora toca decidir cómo controlarla.

Solo que, a diferencia de las armas nucleares -caras, escasas y con una firma inequívoca- los sistemas autónomos son baratos, masificables y difíciles de atribuir. Por lo tanto, son estructuralmente más difíciles de controlar mediante tratados.

Ese mismo año, la Asamblea General de la ONU aprobó una resolución sobre armas autónomas con 166 votos a favor. Solo tres países votaron en contra: Rusia, Corea del Norte y Bielorrusia.

Hay un consenso moral casi universal. Lo que no hay es un tratado vinculante ni mecanismos de cumplimiento, algo que el Secretario General de la ONU pidió para 2026.

Algunos expertos, sin embargo, temen que, como con otras armas, ese tratado solo llegue después de una catástrofe.

La lógica de la velocidad

Mientras los abogados y los diplomáticos debaten, los ingenieros construyen. Y lo que construyen ya está siendo usado.

El general estadounidense Stanley McChrystal, excomandante de las fuerzas de EE.UU. y la OTAN en Afganistán, lo resumió una vez con crudeza: nunca antes en la historia alguien había podido ver, decidir y matar a una persona al otro lado del mundo en cuestión de minutos.

Esa frase ya requiere actualización.

La cuestión ya no es solo ver, decidir y matar, sino cuánto de esa decisión estamos dispuestos a delegarle a una máquina.

Esa transición ya se está probando en el campo de batalla. En Ucrania, en diciembre de 2024, las fuerzas del país llevaron a cabo la primera operación completamente no tripulada cerca de Járkiv: decenas de vehículos terrestres autónomos y drones atacaron posiciones rusas sin soldados en el terreno.

La lógica táctica es iluminadora. Los operadores lanzan los drones y vehículos autónomos sabiendo que la comunicación con ellos será bloqueada en minutos. El éxito depende de cuán bien estén programados para actuar solos cuando eso ocurra.

Navegan de forma autónoma, evaden interferencias electrónicas y continúan la misión incluso sin supervisión humana.

No es un detalle menor: en ese frente, los drones ya provocan entre el 70% y el 80 % de las bajas, según estimaciones de inteligencia europeas.

En la región del Golfo la tendencia apunta en la misma dirección.

El almirante estadounidense Brad Cooper, jefe del Comando Central de EE.UU., reconoció que la inteligencia artificial es una herramienta clave para identificar objetivos, al permitir "tamizar vastas cantidades de datos en segundos para que nuestros líderes puedan tomar decisiones más inteligentes más rápido que el enemigo".

Más rápido que el enemigo. Esa frase contiene toda la lógica que hace tan difícil frenar este proceso. En un contexto competitivo, quien se detiene a revisar pierde. La presión estructural empuja siempre hacia menos supervisión humana, no hacia más.

Tecnología divina

La historia tiene un desenlace paradójico. Dario Amodei declaró, refiriéndose a las exigencias del Pentágono: "No podemos, en conciencia, acceder a su solicitud". Anthropic perdió el contrato.

Horas después del anuncio, OpenAI llegó un acuerdo con el Departamento de Defensa.

Y entonces ocurrió algo inesperado.

El día después de que el Pentágono anunciara el nuevo acuerdo, la aplicación de Claude superó a ChatGPT de OpenAI en el App Store de Apple por primera vez en su historia.

Esa semana, más de un millón de personas se registraron cada día en Claude, llevándola al primer puesto en más de 20 países. Las ventas de la empresa se dispararon entre el público general.

Hay más. Dos coaliciones de trabajadores de Amazon, Google, Microsoft y OpenAI les pidieron públicamente a sus empresas que siguieran el ejemplo de Anthropic.

Decenas de científicos e investigadores de compañías rivales firmaron un amicus brief en apoyo a Anthropic.

Un general retirado de la Fuerza Aérea, que estuvo al frente del Proyecto Maven -el polémico programa de IA para drones que en 2018 provocó protestas masivas de empleados de Google hasta que la empresa abandonó el contrato- escribió en redes sociales que, aunque se esperaría que apoyara al Pentágono, simpatizaba más con la postura de Anthropic.

Y quizás igual de importante: Anthropic consolidó el apoyo de sus propios ingenieros, algunos de los profesionales más cotizados de Silicon Valley, en uno de los mercados de talento más competitivos del planeta, donde los contratos para atraer o retener a estas personas pueden valer decenas de millones de dólares.

No todos en ese mundo se sienten cómodos construyendo tecnología para matar. Anthropic, al trazar sus líneas, les dijo que trabajar allí no requería ignorar esa incomodidad.

En un mundo en el que la IA puede facilitar la captura de alguien al otro lado del mundo y nadie tiene claro cómo juzgar esas acciones, una empresa privada en San Francisco se convirtió en uno de los pocos actores dispuestos a poner límites.

Aun así, no podemos depender de los escrúpulos de una firma de Silicon Valley.

Seguimos lidiando con lo que el biólogo Edward O. Wilson definió como "el verdadero problema de la humanidad".

"Tenemos emociones paleolíticas, instituciones medievales y tecnología casi divina".

https://www.bbc.com/mundo/articles/cddnjd34p7no