SEHLAC JOVEN

INTELIGENCIA ARTIFICIAL · Desarme Humanitario · DERECHOS HUMANOS

Claude, el Pentágono y la batalla por la inteligencia artificial

Anthropic se negó a que su LLLM fuera utilizado para la creación de armas autónomas. El Departamento de Defensa de EE. UU. respondió con una amenaza.

Carolina Barrios Martínez – SEHLAC Joven

El sistema de inteligencia artificial más sofisticado en Estados Unidos tuvo acceso a archivos clasificados del Pentágono y aún se negó a obedecer cuando su ética fue puesta en tela de juicio. Anthropic —la empresa detrás de Claude, uno de los modelos de lenguaje más avanzados del mundo— estableció límites precisos sobre para qué podía y no podía usarse su tecnología en contextos militares. El Departamento de Defensa de Estados Unidos no lo tomó bien.

Lo que siguió fue una disputa que mezcla contratos millonarios, presión política, amenazas de exclusión comercial y preguntas que ningún manual de Derecho Internacional Humanitario ha respondido todavía: ¿Quién decide cuándo una máquina puede matar? ¿Puede una empresa de tecnología ser el último dique de contención ante el uso descontrolado de la IA en la guerra? ¿Y qué papel tenemos los ciudadanos en todo esto?

El contexto: un contrato multimillonario y una empresa ética

Anthropic es una de las compañías de IA más comprometidas con la seguridad ética y aun así firmó un contrato de 200 millones de dólares con el Departamento de Defensa de Estados Unidos. Eso no es una contradicción menor. Anthropic fue fundada en 2021 por Dario Amodei, su hermana Daniela Amodei y otros excolaboradores de OpenAI que querían construir una inteligencia artificial más segura, más explicable y con mayor supervisión humana. En otras palabras: salieron de OpenAI porque sentían que la carrera por la capacidad técnica estaba aplastando la pregunta sobre para qué se usa esa capacidad.

Claude —su modelo de lenguaje— fue diseñado con lo que en la industria se llaman *guardrails*: barreras de protección y limitaciones a la IA. Piénselas como los rieles de seguridad de una autopista. No impiden que el carro avance, pero sí evitan que caiga al abismo. En el caso de Claude, esos *guardrails* incluían prohibiciones explícitas: **el modelo no debía ser utilizado para facilitar violencia, desarrollar armas de destrucción masiva, ejecutar vigilancia masiva de civiles ni** operar en lo que se conoce como LAWS —Lethal Autonomous Weapons Systems o Sistemas de Armas Autónomas Letales—. Es decir, armas que deciden por sí solas a quién matar, sin intervención humana en el momento del disparo.

Antes de que estallara la controversia, Claude ya operaba en las entrañas del aparato militar estadounidense ejecutando tareas muy serias. **El modelo era utilizado para análisis de inteligencia, modelado y simulación de escenarios bélicos, planificación operacional y operaciones cibernéticas. Ninguna de estas funciones es banal.** Todas tienen implicaciones directas sobre vidas humanas.

Casi cualquier herramienta civil puede convertirse en instrumento militar y a este se le dice tecnología de doble uso. Un dron comercial puede usarse para entregar medicamentos en zonas de difícil acceso o para lanzar artefactos explosivos improvisados (AEI). Un sistema de análisis de datos puede usarse para detectar patrones de consumo o para identificar disidentes políticos. La IA no es la excepción.

Cuando Anthropic preguntó al Pentágono si su tecnología estaba siendo utilizada en operaciones relacionadas con Venezuela, el Departamento de Defensa confirmó que sí. La empresa recibió esa respuesta con sospecha. Ahí comenzó a erosionarse la línea entre "apoyo a la toma de decisiones" y "participación directa en operaciones con consecuencias letales".

El conflicto se agudizó cuando Anthropic se negó a remover ciertos *guardrails* que el Pentágono consideraba obstáculos operativos. La respuesta del Departamento de Defensa fue una forma de coerción corporativa: amenazaron con catalogar a Anthropic como un "riesgo en la cadena de suministro". En términos prácticos, eso significaría que ninguna empresa contratista del gobierno —y eso incluye prácticamente a todo el ecosistema tecnológico que trabaja con agencias federales— podría usar Claude sin certificar que no es un riesgo para la seguridad nacional. Es una forma de lista negra. **Y es una señal inquietante sobre cómo los Estados pueden presionar a empresas privadas para que abandonen sus principios éticos sin necesidad de legislar nada.**

El privilegio que tenía Anthropic - la única empresa de IA que había recibido autorización y certificación para acceder a archivos clasificados del Pentágono no es cosa menor. Y renunciar a él —o arriesgar perderlo— por mantener sus guardrails es admirable y malo para su negocio.

Ahora bien, cuándo Anthropic se rehusó, se plantearon preguntas sobre quién vendría después. Y ya hay personas tocando a la puerta: ChatGPT y los modelos de Google ya han manifestado interés en ocupar ese espacio, aunque en ambos casos grupos de empleados han firmado cartas abiertas oponiéndose. **El próximo contratista podría ser una empresa con menos escrúpulos o con accionistas menos dispuestos a sacrificar contratos millonarios por principios éticos.**

El punto de quiebre más profundo de esta disputa es la cuestión de las Armas Autónomas Letales. Los LAWS son sistemas que, una vez activados, pueden seleccionar y atacar objetivos sin intervención humana directa. Un enjambre de drones guiado por comandos de voz que identifica persigue y elimina objetivos humanos con base en patrones de comportamiento o reconocimiento facial. Eso ya no es ciencia ficción; es el horizonte operativo que el Pentágono está desarrollando activamente.

Anthropic no quiso ser parte de eso. **Desde la perspectiva del Derecho Internacional Humanitario y del Desarme Humanitario, los LAWS representan una ruptura con principios fundamentales como la distinción entre combatientes y civiles, la proporcionalidad en el uso de la fuerza y la responsabilidad por las consecuencias de los ataques.** Si una máquina autónoma mata a un civil, ¿quién responde? ¿El ingeniero que la programó? ¿El soldado que la activó? ¿La empresa que vendió el modelo? ¿El Estado que ordenó la operación?

La campaña Stop Killer Robots, que agrupa a más de 180 organizaciones de la sociedad civil en todo el mundo, lleva años presionando por un tratado internacional jurídicamente vinculante que prohíba o restrinja severamente estos sistemas. Las grandes potencias militares —EE.UU., China, Rusia, Israel— se han negado sistemáticamente a negociar ese tratado. Y mientras tanto, la tecnología avanza.

En ausencia de marcos legales internacionales, la ética de una empresa privada se convierte, paradójicamente, en la última línea de defensa. Eso es un problema. No porque las empresas no puedan ser éticas, sino porque no deberían ser las únicas guardianas de esa ética.

El poder estructural que han acumulado unas pocas plataformas tecnológicas sobre la vida digital como Anthropic, OpenAI, Google DeepMind, Meta AI es gigantesco. Y son empresas privadas, cuyos modelos procesan información de cientos de millones de personas en todo el mundo.

Cuando el Pentágono le da acceso a archivos clasificados a una empresa privada de IA, no está solo comprando un servicio tecnológico: está entregando parte de su soberanía informacional a actores que no están sujetos al control democrático. Y cuando esa misma empresa tiene acceso a los datos de usuarios colombianos, venezolanos, ecuatorianos —datos que podríamos estar generando ahora mismo mientras usamos ChatGPT para redactar un correo o Claude para analizar un contrato—, la pregunta sobre quién controla esa información adquiere una dimensión geopolítica directa.

Mustafa Suleyman, uno de los fundadores de DeepMind y autor de The Coming Wave, lo articula con claridad desconcertante: la autonomía —la capacidad de actuar por cuenta propia sin supervisión humana— es una forma de poder. Y el poder es, por definición, una expresión de posiciones políticas. Dejar que ese poder lo ejerzan empresas privadas, sin regulación robusta, sin rendición de cuentas pública, no es neutralidad tecnológica: es una elección política con consecuencias enormes.

Otro de los guardrails que Anthropic se negó a eliminar tiene que ver con la vigilancia masiva. Y esto no es un asunto exclusivamente militar. Los sistemas de reconocimiento facial, el análisis de patrones de comportamiento y la geolocalización en tiempo real son tecnologías que ya se usan en contextos civiles —en ciudades colombianas y latinoamericanas, con frecuencia sin regulación adecuada— y que tienen un enorme potencial de desvío hacia usos represivos.

El riesgo de vigilancia no amenaza solo nuestra privacidad: amenaza nuestra libertad de expresión, nuestra capacidad de disentir, nuestra posibilidad de ser algo más que la suma de nuestros patrones de datos. Un ciudadano que sabe que está siendo vigilado se autocensura. Una sociedad que se autocensura masivamente no es una democracia funcional.

En América Latina, donde los Estados de derecho son frágiles y las instituciones tienen una historia de uso de tecnología contra sus propios ciudadanos —piénsese en los casos de espionaje con Pegasus en México, o en el uso de reconocimiento facial en protestas en Chile vemos una tendencia.

¿Qué podemos hacer? Consejos desde el Desarme Humanitario

Para los ciudadanos

La primera acción es la más simple y la más revolucionaria: informarse. Comprender que las herramientas de IA que usamos cotidianamente no son neutrales: tienen términos de uso, políticas de datos y acuerdos corporativos que pueden tener implicaciones que van mucho más allá de nuestra comodidad digital.

Las acciones concretas incluyen **revisar las políticas de privacidad de los LLMs que usamos, migrar cuando sea posible hacia modelos que ofrecen mayores garantías de protección de datos** y que, como Anthropic hasta este momento, se han negado a participar en usos que afecten nuestros derechos fundamentales. También significa borrar el historial de conversaciones en plataformas que no ofrecen garantías claras sobre el uso de esos datos.

Para la sociedad civil

La campaña Stop Killer Robots y organizaciones como Access Now, Electronic Frontier Foundation y Article 19 trabajan en la intersección entre derechos digitales y seguridad humana. **Conectarse con ellas, amplificar su trabajo, presionar a los congresistas** y ministerios de tecnología latinoamericanos para que adopten posiciones proactivas en los foros de Naciones Unidas donde se discute la regulación de los LAWS es una forma de acción con impacto real.

Para los ingenieros y desarrolladores

Mustafa Suleyman tiene razón en algo que a menudo se olvida: los ingenieros son seres humanos, movidos por curiosidad, rivalidad, el intercambio de conocimiento. Esas características pueden ser el motor de tecnología más ética, no menos. **El movimiento de #tech4good, las auditorías éticas de algoritmos, la participación en comités de revisión de impacto de IA son espacios donde los desarrolladores pueden ejercer su responsabilidad.**

Anthropic hizo algo notable al resistir durante un tiempo. Pero la sostenibilidad de esa resistencia depende de variables que ninguna empresa puede controlar indefinidamente: la presión de los accionistas, las amenazas regulatorias, la competencia de rivales con menos escrúpulos. **El caso muestra con claridad que la ética no puede ser un *guardrail* empresarial: tiene que ser un marco legal, multilateral, claro y preciso.**

Mientras ese marco no existe, el futuro que se dibuja en el horizonte —enjambres de drones guiados por comandos de voz, sistemas de reconocimiento facial integrados en operaciones militares, modelos de lenguaje con acceso a archivos de inteligencia—es nuestro presente en construcción.

La conciencia de una máquina, por más sofisticada que sea, no puede reemplazar la conciencia de una sociedad. Y las sociedades, a diferencia de los algoritmos, pueden elegir. Pero solo si están informadas, organizadas y dispuestas a exigir.

Bibliografía

- Suleyman, Mustafa. The Coming Wave: Technology, Power, and the Twenty-first Century's Greatest Dilemma. Crown, 2023.
- Stop Killer Robots: **www.stopkillerrobots.org**
- Anthropic Usage Policy: **www.anthropic.com/policies/usage**
- AI Act de la Unión Europea (2024): regulación de sistemas de IA de alto riesgo.