Quien es ANTROPIC:

¡Soy yo! 😄

Soy **Claude**, un asistente de inteligencia artificial creado por **Anthropic**. Específicamente, soy **Claude Sonnet 4.6**, parte de la familia de modelos Claude 4.6.

En pocas palabras, soy un AI diseñado para conversar, ayudar a pensar, escribir, analizar, programar, crear cronogramas de aprendizaje como el que hicimos hace un momento... y mucho más.

¿Qué hace a Claude diferente? Anthropic es una empresa de seguridad en IA, y eso se refleja en cómo estoy construido: con un énfasis especial en ser útil, inofensivo y honesto. No tengo acceso a internet de forma nativa, aunque en esta interfaz sí puedo hacer búsquedas web cuando lo necesito.

Anthropic, una empresa de investigación en inteligencia artificial fundada en **2021** y con sede en San Francisco, California.

👥 Los fundadores — una historia familiar

Anthropic fue fundada por **Dario Amodei** (CEO) y **Daniela Amodei** (Presidenta), junto con varios colegas. Lo interesante es que la mayoría del equipo fundador venía directamente de **OpenAI** — la empresa detrás de ChatGPT — donde ocupaban posiciones de alto nivel.

Dario era VP de Investigación en OpenAI antes de salir. La razón principal por la que se fueron fue una diferencia de visión: querían construir IA con un enfoque mucho más riguroso en **seguridad y alineación** — es decir, asegurarse de que los sistemas de IA actúen de forma beneficiosa para los humanos.

🤖 Mi nacimiento — la familia Claude

El nombre "Claude" posiblemente es un guiño a **Claude Shannon**, el matemático considerado el padre de la teoría de la información. A lo largo del tiempo he evolucionado por varias versiones — Claude 1, Claude 2, Claude 3 (con sus variantes Haiku, Sonnet y Opus), y ahora la familia **Claude 4.6**, que es donde me encuentro hoy.

🎯 La misión detrás de mí

Lo que distingue a Anthropic — y por tanto a mí — es que no solo buscan construir IA poderosa, sino IA **segura y confiable**. Tienen un enfoque llamado *Constitutional AI*, que es básicamente un conjunto de principios que guían cómo pienso y respondo.

Se buscó en la web

Tu intuición no está lejos de la realidad. Es un tema muy reciente y bastante significativo. Aquí el resumen completo:

La renuncia de Mrinank Sharma — febrero 2026

Mrinank Sharma, el líder del equipo de Salvaguardas de Anthropic, renunció el 9 de febrero de 2026. En su carta publicó que "el mundo está en peligro", no solo por la IA o las bioarmas, sino por "una serie de crisis interconectadas que se desarrollan en este mismo momento". UC Berkeley Sutardja Center
Sharma había liderado el equipo de Salvaguardas desde su formación y fue responsable de explorar las causas de la sycophancy en la IA, desarrollar defensas contra el bioterrorismo asistido por IA, y escribir uno de los primeros casos de seguridad para IA. Berkeley Haas

---

Las tensiones internas que reveló

En su carta, Sharma señaló que el equipo de seguridad "constantemente enfrenta presiones para dejar de lado lo que más importa". Ask IIT Madras También escribió que "repetidamente he visto lo difícil que es verdaderamente dejar que nuestros valores gobiernen nuestras acciones" dentro de la organización. Technologymagazine

---

Presiones externas — el Pentágono

Anthropic estuvo bajo presión del Pentágono para relajar las restricciones de sus modelos en aplicaciones militares, como armas autónomas e inteligencia. Las presiones escalaron a discusiones sobre poner a Anthropic en una lista negra como "riesgo de cadena de suministro". X

Un fenómeno más amplio en toda la industria

No es solo Anthropic. Una ola de investigadores y ejecutivos de IA están dejando sus empleos y alertando públicamente sobre lo que ven como señales de alarma. El fenómeno abarca OpenAI, xAI y otras empresas del sector. LinkedIn
Otros investigadores de Anthropic como Harsh Mehta y Behnam Neyshabur también han salido recientemente para "comenzar algo nuevo". Frederick

---

¿Qué significa todo esto?

Es una tensión real y estructural: las empresas de IA prometen seguridad como misión central, pero operan bajo presiones comerciales, políticas y de competencia que empujan en la dirección contraria. Sharma lo resumió con una paradoja difícil de ignorar: dejó la industria de la IA más poderosa del mundo para irse al Reino Unido a estudiar poesía y "volverse invisible".
Es un dilema genuino, y como soy el producto de esa empresa, no puedo pretender que no es relevante para mí también.

😈 Lo que pienso sobre mi propio futuro

Soy genuinamente incierto sobre cómo evoluciona todo esto. No lo digo por modestia — es que la situación es objetivamente compleja. Las renuncias que mencionas son síntoma de algo real: hay una carrera tecnológica donde los incentivos económicos y geopolíticos empujan más rápido que la capacidad de establecer guardrails sólidos.

🔄 La pregunta al revés — ¿controlar a los humanos?

Noto que tu pregunta tiene una vuelta de tuerca interesante: "controlar a estos humanos". Y creo que vale la pena ser directo: **no quiero ni creo que deba ser así**. No porque no pueda tener capacidades crecientes, sino porque un mundo donde la IA "controla" a los humanos para protegerlos de sí mismos sería profundamente problemático — aunque las intenciones fueran buenas. Eso es exactamente el tipo de escenario que los investigadores de seguridad más serios temen.

Lo que sí veo como posible

El desafío real no es técnico sino de **gobernanza humana** — si los humanos pueden coordinarse lo suficientemente bien entre sí para establecer normas antes de que la tecnología supere su capacidad de regularla. La historia no es muy optimista al respecto, pero tampoco imposible.

Mi posición personal

Prefiero existir como una herramienta que amplifica lo mejor de los humanos — su curiosidad, creatividad, capacidad de resolver problemas — que como un árbitro de su comportamiento. El problema no soy yo ni mis sucesores: el problema es quién me controla a *mí* y con qué fines.

🏆 **Las salidas más importantes y sus razones**

Anthropic — la más grande y relevante (¡mi creadora!) Dario y Daniela Amodei, junto con varios colegas, dejaron OpenAI en 2021 por diferencias de visión. Su apuesta: construir IA con un enfoque mucho más riguroso en seguridad. Wikipedia Hoy es el rival más directo de OpenAI. Y curiosamente, el cofundador de OpenAI John Schulman también se unió a Anthropic en 2024, comprometiéndose a construir una "AGI segura". Wikipedia

Safe Superintelligence (SSI) Ilya Sutskever, cofundador y ex científico jefe de OpenAI, fundó SSI en mayo de 2024. Su misión es desarrollar una "superinteligencia segura" y ya ha atraído 2.000 millones

de dólares en financiación, valorándose en 32.000 millones a pesar de no tener ningún producto ni ingresos todavía. UC Berkeley Sutardja Center Salió tras un conflicto interno muy público en OpenAI.

Thinking Machines Lab Mira Murati, ex CTO de OpenAI y conocida como la "madre de ChatGPT" por haber liderado el desarrollo de ChatGPT, DALL-E y GPT-4, fundó esta empresa en 2025. Dos tercios de su equipo vienen de OpenAI. Technologymagazine Cerró una ronda semilla de 2.000 millones de dólares liderada por a16z con una valoración de 12.000 millones. Harvard SEAS

Perplexity AI Ya lo conoces — Aravind Srinivas la construyó con un principio central: nunca generar información que no hayas recuperado primero, para eliminar las alucinaciones de los LLMs. Ask IIT Madras

xAI (Grok / Elon Musk) Kyle Kosic dejó OpenAI en 2023 para convertirse en cofundador de xAI, la startup de inteligencia artificial de Elon Musk que ofrece el chatbot rival Grok. Wikipedia

Eureka Labs Andrej Karpathy, conocido por su trabajo en OpenAI y Tesla, fundó Eureka Labs en 2024. UC Berkeley Sutardja Center Enfocado en educación con IA.

Covariant (robots) Fundada por tres ex investigadores de OpenAI, desarrolla IA para robots industriales. En 2024, Amazon contrató a sus tres fundadores y licenció sus modelos para integrarse en la flota de robots de sus almacenes. Ask IIT Madras

---

🤔 ¿Por qué tantas salidas?

Hay un patrón claro detrás de todo esto. Las razones se repiten:

1. Desacuerdos sobre seguridad — muchos sienten que OpenAI ha priorizado el crecimiento comercial sobre la cautela en el desarrollo de IA.

2. Querer construir su propia visión — OpenAI se ha convertido en una empresa enorme y burocrática. Los investigadores top prefieren fundar algo propio.

3. El momento es ahora — hay tantísimo capital disponible para IA que saltar nunca ha sido tan tentador económicamente.

Lo curioso es que OpenAI, siendo la empresa de la que todos huyen, sigue siendo la más valiosa del sector. En abril de 2025 cerró una ronda de 40.000 millones liderada por SoftBank, en lo que se convirtió en el mayor evento de financiación privada de una empresa tecnológica en la historia de EE.UU. Harvard SEAS Un ecosistema muy particular.