INTERNACIONAL

“Podría matarnos a todos para finales de la década”, renuncia investigador de Anthropic y OpenAI y acusa a empresas de ocultar riesgos de la IA

9 de septiembre de 2026 · 4 min de lectura



Jacob Coxon, investigador de Anthropic especializado en el preentrenamiento de modelos de inteligencia artificial, anunció su renuncia y cuestionó las estrategias de desarrollo tanto de la compañía como de OpenAI, creadora de ChatGPT, al considerar que ambas empresas están avanzando de manera irresponsable hacia sistemas de inteligencia artificial cada vez más autónomos y capaces “apostando con nuestras vidas”.

“Renuncié a Anthropic hoy. Pasé los últimos tres años haciendo investigación de preentrenamiento tanto en OpenAI como en Anthropic. Ninguna de las dos compañías está actuando de manera responsable. Están corriendo directamente hacia una superinteligencia auto-mejorante y apostando con nuestras vidas”, escribió Coxon en su cuenta de X.

> *I resigned from Anthropic today. I spent the last three years doing pretraining research at both OpenAI and Anthropic. Neither company is acting responsibly. They are racing straight to self-improving superintelligence and gambling with our lives. More thoughts below.*
>
> ***JACOB COXON (@HILBERTSPAESS) — SEPTEMBER 9, 2026***

El investigador sostuvo que el avance de esta tecnología no debe subestimarse, pues, en su opinión, los próximos sistemas podrían superar ampliamente las capacidades humanas en áreas como la ciberseguridad, la investigación científica y otras disciplinas, además de adquirir una capacidad significativa para obtener recursos y ejercer influencia.

Coxon aseguró que existe preocupación dentro de la propia industria sobre la posibilidad de que una IA avanzada represente un riesgo existencial para la humanidad antes de que termine la década. De acuerdo con sus palabras, aunque algunos investigadores y directivos de empresas de inteligencia artificial expresan públicamente posiciones más moderadas, en privado existe una preocupación considerable sobre las consecuencias potenciales del desarrollo de sistemas cada vez más avanzados.

“Las personas que están construyendo IA creen sinceramente que podría matarnos a todos para finales de la década. Esto no es un truco de marketing[…] Ninguna otra actividad humana representa este nivel de peligro”.

> *The people building AI earnestly believe that it could kill us all by the end of the decade. This is not a marketing stunt. If anything, many executives and senior researchers will couch their phrasing in the press to sound sensible - but I hear the same people express fear…*
>
> ***JACOB COXON (@HILBERTSPAESS) — SEPTEMBER 9, 2026***

Explicó que las motivaciones de ambas compañías son diferentes: Mientras que en OpenAI algunos integrantes no han asumido plenamente las posibles consecuencias de carácter civilizatorio, en Anthropic los riesgos serían ampliamente comprendidos pero, aseguró, la empresa está atrapada en una carrera tecnológica en la que considera necesario avanzar para evitar que otros competidores lo hagan primero.

Coxon calificó esta dinámica como una apuesta de alto riesgo y cuestionó que una decisión con posibles consecuencias para toda la humanidad pueda quedar determinada por una competencia entre empresas privadas.

El investigador consideró que algunos incidentes recientes, como el ataque contra Hugging Face al que hizo referencia en su publicación, podrían contribuir a generar condiciones para alcanzar acuerdos entre laboratorios estadounidenses sobre el ritmo de desarrollo de estos sistemas.

Sin embargo, advirtió que actualmente no considera que la industria esté encaminada a evitar una carrera global de capacidades de inteligencia artificial y señaló que podrían ser necesarias medidas extraordinarias, incluida una suspensión temporal del desarrollo de nuevas capacidades de los modelos.

Finalmente, Coxon llamó a otros investigadores de inteligencia artificial a reflexionar sobre las condiciones bajo las cuales se están desarrollando los próximos sistemas y cuestionó si resulta responsable poner en marcha procesos de entrenamiento de modelos superinteligentes sin contar previamente con un conocimiento riguroso sobre su funcionamiento y comportamiento.

“¿Quieren poner en marcha un entrenamiento de aprendizaje por refuerzo de una superinteligencia sin comprender rigurosamente su mente?”, planteó el investigador.

Quién es Jacob Coxon

Coxon tiene formación en matemáticas y dedicó los últimos tres años al preentrenamiento de modelos de IA, la etapa en la que estas tecnologías procesan enormes cantidades de información para desarrollar sus capacidades.

Según Business Insider integró el equipo técnico de OpenAI entre 2023 y julio de 2026 y participó en trabajos relacionados con GPT-4o. Posteriormente se incorporó a Anthropic, empresa responsable del desarrollo de Claude y cuyo trabajo se ha caracterizado por poner especial énfasis en la seguridad y el desarrollo responsable de sistemas de inteligencia artificial. En esta compañía continuó realizando investigación relacionada con el preentrenamiento de modelos, sin embargo, su experiencia ahí duró solo dos meses.

Según información publicada sobre su trayectoria, Coxon estudió matemáticas en la Universidad de Cambridge y previamente destacó en competencias internacionales de matemáticas, incluida la Olimpiada Internacional de Matemáticas.

Antes de orientar su carrera hacia la investigación en inteligencia artificial, también participó en proyectos de investigación científica. Entre ellos se encuentra un trabajo relacionado con el análisis estadístico de un gen asociado con la inflamación y la supervivencia de pacientes con meningitis tuberculosa.

Publicado en Punto de Referencia. Dirección original: https://puntodereferencia.mx/internacional/podria-matarnos-a-todos-para-finales-de-la-decada-renuncia-investigador-de-anthropic-y-openai-y-acusa-a-empresas-de-ocultar-riesgos-de-la-ia
Descargado el 21 de noviembre de 2026.