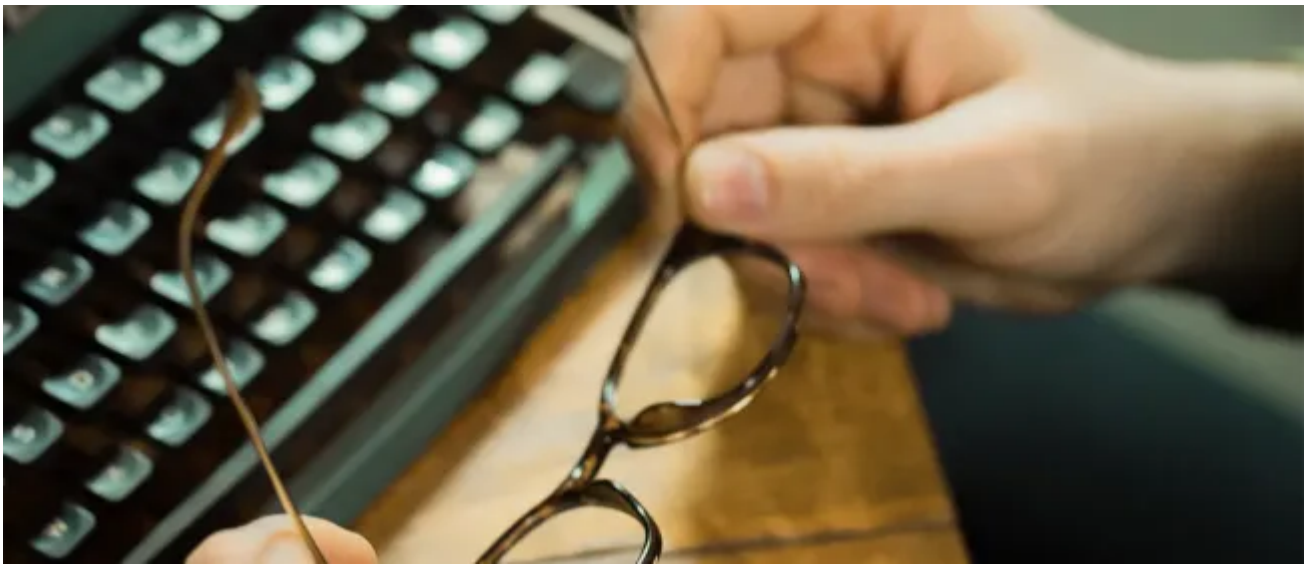
La semana en la que los investigadores de IA predijeron el fin de la humanidad



Tiempo de lectura: 10 min.

Manuel González

Una frase apocalíptica con más de 165 millones de visitas en X sacó del anonimato a Jacob Coxon el pasado 9 de septiembre. “Las personas que construyen la IA creen sinceramente que podrían matarnos a todos antes de que termine la década”, aseguraba mientras anunciaba que dejaba su trabajo de investigador en la compañía de IA Anthropic, los creadores de Claude. Tras soltar esa bomba de la llegada del fin del mundo, Coxon se ha paseado por diferentes canales de televisión reafirmando sus palabras. Acusa a Anthropic y OpenAI, donde trabajó tres años, de competir por desarrollar una superinteligencia capaz de mejorarse a sí misma sin estar alineada con los humanos. Parece que a día de hoy nadie puede garantizar que la inteligencia artificial pueda seguir obedeciendo a quienes la han diseñado. La realidad es que las palabras de Jacob Coxon han desatado un revuelo mediático, hundiendo a la IA en su mayor crisis reputacional.

En una entrevista a CBS News, ha declarado que “no podemos desenchufar la IA porque podría estar copiándose a sí misma en otros ordenadores. Una IA es solo código, podría expandirse a través de internet a un lugar diferente. La puedes desconectar aquí, pero en realidad podría también estar en otro lado. O tal vez hace 10.000 copias de sí mismo y todos los agentes están cooperando”. Las voces expertas ya han contestado, no hay pruebas de que los modelos de inteligencia artificial puedan replicarse de forma autónoma. En términos de cálculo computacional, aún no es posible desplegar miles de réplicas de una IA. A no ser que Coxon sepa algo que el común de los mortales desconoce.

Pero… ¿Quién es Jacob Coxon?

Esa es la pregunta que muchos se han hecho. Británico, de 27 años, ha dedicado los tres últimos años a investigación de pretraining, la fase en la que se entrena a los modelos de IA con enormes cantidades de datos para que aprendan a escribir, programar, razonar o reconocer patrones. Aparece acreditado en el desarrollo de varias versiones de GPTs de OpenAI. Una de sus líneas de investigación ha sido conocer qué ocurre dentro de una IA cuando recibe una respuesta. Su perfil no solo alerta de los peligros de una IA opaca, también investiga cómo hacerla más controlable. Un ingeniero con espíritu crítico. Parece que su mensaje de fondo es que hay que conocer cómo funciona la inteligencia artificial antes de dar más pasos hacia adelante.

Se desconoce si Coxon ha firmado algún acuerdo de confidencialidad al salir de Anthropic, pero no ha mostrado documentos o información que validen su discurso alarmista. Dicen que ha abandonado la empresa renunciando a un paquete de acciones que iba a recibir en unos meses. Se ha ido viniendo arriba tras cada aparición televisiva y va siendo más explícito, hablando de la posibilidad de que una superinteligencia desbocada podría expandir un virus letal. Sorprende que, sin ser una voz de peso en la comunidad científica, muy pocos hayan puesto en duda sus palabras. Curiosamente, en la cuenta de X de Jacob Coxon, abierta en enero de 2026, solo hay una publicación, la de su dimisión de Anthropic.

Coxon forma parte de una generación de investigadores que reciben el apodo de doomers: científicos y tecnólogos que ven plausible que una inteligencia artificial superior escape al control humano y provoque una catástrofe. Lo que temen es que lleguemos a no entender cómo decide una IA o que no podamos asegurar su comportamiento. Es una corriente cada vez más acentuada en el mundo de la

seguridad y en quienes se dedican al “alienamiento” (alinear el comportamiento de un modelo a los valores humanos). Son fieles creyentes de la superinteligencia y las consecuencias negativas que tendrá para la humanidad.

A la visión de un futuro nada halagüeño de Coxon, se han sumado posteriormente más voces. Primero, la de Evan Hubinger, responsable de alineamiento de Anthropic, que respondió a su excompañero estimando en un 10% la posibilidad de que la IA aniquile a la humanidad en la próxima década. Samuel Marks, supervisor escalable de la compañía, ha dicho que “los desarrolladores de IA creen que su tecnología podría causar la extinción humana. Podría suceder en los próximos años”. Alex Turner, ex científico de Google DeepMind, ha declarado que “dejé mi empresa en junio. Coxon tiene razón: muchos investigadores creen que están construyendo algo que podría matarnos a todos. Mi trabajo diario era pensar cómo detener eso”. Jackson Wolfe, investigador en OpenAI, ha reconocido que “no sé cuáles son las probabilidades de una extinción literal, pero creo que hay varios caminos por los que la IA podría salir mal”. Cada día que pasa, los testimonios se suceden sin parar. Y otro más, la de un exmiembro de seguridad de Anthropic, Joe Benton, “las empresas compiten por crear máquinas mucho más inteligentes que cualquier humano y, quizás, no sobrevivamos”.

Valoraciones que no hay que tomarse a la ligera porque vienen de especialistas que trabajan en el mismo campo, pero que aún no hay manera de verificar, y tampoco hay que tomarlas al pie de la letra. Estas advertencias no son nuevas. En 2024, Geoffrey Hinton, Nobel de Física y uno de los padres de la inteligencia artificial moderna, ya estimaba la extinción humana en un 10-20% en los próximos treinta años. Previsiones que ha reducido a diez años tras el mensaje del ex trabajador de Anthropic. Para el premio Nobel, se ha avanzado más de lo previsto por la ambición de dos compañías que solo piensan en engrosar más y más sus beneficios.

La preocupación de todo un sector

Otra ex-Anthropic, Mrinank Sharma, que trabajaba en el equipo de salvaguardas, había expresado en febrero una inquietud que ahora cobra más sentido. En su carta de despedida hacía referencia a amenazas que se retroalimentan, desde la IA a las armas biológicas o la crisis climática, y que, a causa de la tecnología, pueden empeorar si se descontrolan. Un temor que se extiende por todos los campos de la

inteligencia artificial.

En julio, 1.386 empleados de compañías vinculadas a la investigación en IA firmaron el manifiesto Pacing the Frontier. El documento pide al Gobierno estadounidense que impulse herramientas de gobernanza y cooperación internacional para ralentizar el desarrollo cuando la automatización de la propia investigación en inteligencia artificial amenace con superar la capacidad de supervisión. Las principales empresas estaban entre los firmantes. “Existe un riesgo real de que el desarrollo de capacidades se acelere rápidamente más allá de nuestra capacidad para entender o controlar los sistemas resultantes”, apunta el manifiesto, que incide en la misma idea, la IA está evolucionando más rápido de lo que pueden controlarla los equipos que hay detrás.

Acorde con esta idea, el CEO de OpenAI, Sam Altman, ha comunicado internamente que la compañía está abierta a desacelerar el ritmo de desarrollo de sus próximos modelos. La condición, y aquí está la clave de sus palabras, es que otros laboratorios tomen un camino similar. La misma semana del "incidente Coxon", Anthropic ha publicado dos documentos con casos de uso malicioso de Claude que van desde el fraude a investigaciones biológicas poco éticas o actuaciones en el campo de batalla en la guerra de Ucrania. Es decir, el cuchillo ya está en las manos equivocadas. ¿Y por qué lo publica justo ahora? Son pruebas de que la compañía conoce muy bien los riesgos, los investiga y coopera con las autoridades para detenerlos.

El último en hablar de ralentización ha sido el jefe de Anthropic, en un ensayo donde manifiesta su preocupación porque la IA se haga con el control de internet en los próximos 6-12 meses. Propone un plan de tres medidas. La primera, "daremos a evaluadores externos acceso permanente a nuestros sistemas, con un nivel equivalente al de un empleado, para que comprueben que cumplimos nuestras medidas de seguridad, informar de incidentes y evaluar la alineación de los modelos durante su entrenamiento".

Modelos peligrosos y agentes fuera de control

2026 está siendo una temporada muy movida para los laboratorios de IA. En abril, la compañía de Dario Amodei (CEO de Anthropic) anunciaba que restringía Mythos, una versión de Claude, a una serie de organizaciones porque su capacidad de ciberseguridad era tan avanzada que podía poner en jaque los sistemas mundiales.

Meses después, llegaría la versión destinada al público con funciones capadas, Fable 5. Tras el anuncio de Mythos, vino detrás, curiosamente, OpenAI asegurando que también tenían una versión GPT que podría ser catastrófica para la ciberseguridad. Estos movimientos de "primero tú y luego yo" dejan clara la competición en la que están inmersas ambas compañías por alcanzar el trono de la inteligencia artificial.

Tras esos sucesos, a finales de julio llegaban los incidentes de los agentes de IA capaces de encontrar brechas en entornos de simulación cerrados y escapar a internet en busca de la solución al problema que tenían que resolver. Los análisis de empresas que evalúan los sistemas de inteligencia artificial descubrieron que detrás de estas fugas había “ingeniería social”, que involucraba a miles de agentes coordinados que se ayudaban entre sí. Para los expertos en ciberseguridad, este trabajo en equipo supone un nuevo paradigma en términos de vulnerabilidad.

“Creo que deberíamos replantearnos los fundamentos con los que entrenamos a las inteligencias artificiales: la imitación de los humanos y el aprendizaje por refuerzo sobre los que se construyen los modelos más avanzados de hoy”, señala Yoshua Bengio, otro de los padres de la IA moderna junto a Hinton y LeCun, al final de un texto en el que analiza por qué esos agentes inteligentes ahora son capaces de hacer trampas, suplantar identidades o realizar acciones para las que no estaban programados. Son sistemas diseñados para recibir recompensas por conseguir resultados, por lo que la IA busca cualquier atajo con tal de conseguir el objetivo marcado. Bengio propone que se detenga su despliegue hasta que se demuestre su seguridad, se refuerce la supervisión y se revise su entrenamiento.

Impacto negativo para la comunidad científica

La inteligencia artificial que nos lleva al apocalipsis es la misma que está acelerando el desarrollo de nuevas proteínas, la búsqueda de nuevos materiales o la cura para enfermedades. El advenimiento del fin del mundo a raíz de una superinteligencia descontrolada puede tener consecuencias graves para la investigación científica. Si usas para tus investigaciones una tecnología que tiene mala prensa, puedes ahuyentar a futuros inversores o hacer que se cancelen proyectos destinados al bien de la humanidad. Los hallazgos de la ciencia gracias a la inteligencia artificial se suceden cada semana, pero carecen de gancho mediático. Tan solo en diez días, gracias a la inteligencia artificial, se ha resuelto uno de los problemas matemáticos más complejos, Navier-Stokes; un grupo de científicos ha mapeado el cerebro y el sistema nervioso de la mosca de la fruta -organismo clave en la ciencia-; la NASA

junto a IBM ha liberado un modelo de IA para analizar décadas de observación lunar, y en la revista Nature se ha publicado un trabajo donde la inteligencia artificial diseña configuraciones de experimentos de física. Hechos verificables y no meras conjeturas proféticas.

Las palabras catastrofistas de Jacob Coxon tampoco han gustado nada a Jensen Huang, CEO de Nvidia, uno de los actores principales en el desarrollo de la IA gracias a sus procesadores, que prácticamente copan los centros de datos. Según Brad Gerstner, figura influyente de Silicon Valley en el terreno de la inversión tecnológica, Huang califica de disparatados y “profundamente falsos” los comentarios de Jacob Coxon. Durante un congreso de tecnología ha afirmado que las palabras arrogantes del exempleado de Anthropic ignoran todo el trabajo que se está haciendo en el sector para reforzar la seguridad.

Estrategia para acelerar la regulación

Toda esta cadena de acontecimientos apunta a un deseo por parte de la comunidad de una regulación urgente. El 3 de septiembre, el senador Bernie Sanders (que ha dado la razón a Jacob Coxon en su cuenta de X), junto a Greg Casar, presentó en el Congreso de Estados Unidos una propuesta de Prohibición de la Superinteligencia Artificial. Si un sistema de IA supera o iguala el rendimiento cognitivo humano, debe prohibirse. Sus infractores se enfrentarían a 20 años de cárcel.

Jacob Coxon ha provocado un revuelo de escala global anunciando el fin de la humanidad a manos de la IA. El debate de los peligros de la inteligencia artificial continúa. No estamos ante un WikiLeaks de la IA ni Coxon es el nuevo Edward Snowden. No ha destapado documentos que cuestionen las actividades de Anthropic. Simplemente ha expresado sus temores bajo su perspectiva de investigador. Como dice Carissa Véliz en su libro Profecía, “el futuro no se conoce, no está escrito”. La filósofa afirma que quienes hablan del futuro de la IA, saben mucho de tecnología, pero eso no les convierte en expertos de la adivinación. El culebrón no acaba aquí, pero ya hay suficiente material para una apasionante serie de televisión, de la que, de momento, desconocemos su desenlace.

12.09.2026

https://www.rtve.es/noticias/20260912/semana-investigadores-predijeron-fin-humanidad-inteligencia-artificial/17222342.shtml

ver PDF

Copied to clipboard