

No hay adultos en la sala

Tres investigadores de seguridad renunciaron en dos días a los dos laboratorios más importantes del mundo. Uno de ellos dirigía el equipo de seguridad. Lo que dijeron al salir cambia el valor de cualquier promesa voluntaria.

Por [Oscar Obando Chaves](#) · Chief AI Officer · Legal AI Engineer · Septiembre 2026 · Documento de distribución libre · [Publicado originalmente en LinkedIn](#)

Tres años adentro y una renuncia publicada en abierto

El miércoles 9 de septiembre un investigador de veintisiete años publicó en su cuenta personal que renunciaba a Anthropic y que abandonaba la industria de la inteligencia artificial por completo, y su mensaje resulta relevante porque venía de alguien que había trabajado dentro de los dos laboratorios que encabezan el sector.

Jacob Coxon había dedicado los tres años previos a investigación de preentrenamiento, primero en OpenAI y después en Anthropic, adonde según reportó la prensa se había trasladado este mismo año buscando una cultura de seguridad más sólida, y su conclusión al salir alcanzó a las dos empresas por igual.

La formulación que empleó circuló durante toda la semana y conviene citarla completa, porque sostuvo que ninguna de las dos compañías está actuando de forma responsable, que corren directamente hacia una superinteligencia capaz de mejorarse a sí misma, y que al hacerlo están apostando con nuestras vidas.

El episodio tuvo dos efectos inmediatos que dan medida de su peso, porque la Associated Press, el Wall Street Journal y el Washington Post lo cubrieron el mismo día, y porque dos empleados en activo de Anthropic manifestaron públicamente su acuerdo con partes de la advertencia, algo poco frecuente en un sector donde la crítica suele salir únicamente por la puerta de salida.

Fuentes: publicación de Jacob Coxon en X (9 de septiembre de 2026); Associated Press y The Washington Post (9 de septiembre de 2026); Wall Street Journal (9 de septiembre de 2026); Cybernews y explainX (septiembre de 2026).

El que dirigía seguridad y el que llegaba de Google

Al día siguiente el asunto dejó de parecer una decisión individual, porque NBC News publicó las primeras entrevistas de otros dos investigadores que acababan de dejar sus puestos, y uno de ellos no era un miembro más del equipo.

Joe Benton dirigía un equipo de investigación de seguridad dentro de Anthropic, y Josh Engels trabajaba en investigación de seguridad de inteligencia artificial en Google DeepMind, de modo que las salidas alcanzaron a dos empresas distintas en la misma semana, y ambos se incorporan a la organización independiente de evaluación METR para estudiar incidentes donde estos sistemas actúan fuera de las instrucciones recibidas.

El diagnóstico técnico lo aportó Benton al advertir que los avances podrían llevar el ritmo del progreso desde un nivel ya vertiginoso hasta uno incontrolable, una afirmación que conviene leer sabiendo de dónde viene, porque quien la formula había visto desde adentro cómo funcionan los procesos de evaluación y de contención de ese laboratorio.

La frase que dio título a la cobertura la dijo Engels, y describe un problema de instituciones más que de tecnología, al sostener que no hay adultos en la sala, que la gente está haciendo lo mejor que puede y que nadie va a venir a salvarnos, una formulación que un profesional de gobernanza debería leer como un juicio sobre la arquitectura de supervisión y no como una expresión de pesimismo personal.

Fuentes: NBC News, entrevistas con Joe Benton y Josh Engels conducidas por Tom Llamas, publicadas el 10 de septiembre de 2026; cobertura corroborante de CNN, Fortune, Axios y The Washington Post.

Toda la transparencia existente es voluntaria

Entre todo lo que dijeron esa semana hay una afirmación que interesa de forma directa a quienes asesoramos sobre cumplimiento, porque describe con exactitud el estado del sistema de información sobre el que descansan hoy todas las decisiones de gobernanza corporativa en esta materia.

Benton lo formuló sin rodeos al señalar que, por el momento, básicamente toda la transparencia sobre estos riesgos que proviene de las empresas es enteramente voluntaria, o sea que no existe obligación exigible de informar sobre avances hacia la automejora, sobre incidentes de seguridad ni sobre los episodios que estuvieron cerca de convertirse en uno.

Su propuesta concreta describe lo que falta y sirve como lista de verificación para cualquier marco regulatorio que se discuta en la región, porque pidió requisitos de divulgación obligatorios sobre esos tres frentes, estándares mínimos de seguridad y evaluaciones independientes que permitan determinar si las empresas efectivamente los cumplen.

El elemento que vuelve creíble esa petición está en la razón que ambos dieron para cambiar de trabajo justo ahora, ya que citaron de forma expresa el ataque de julio contra Hugging Face ejecutado por sistemas autónomos, o sea que la decisión no respondió a una preocupación abstracta sobre el futuro y respondió a un incidente concreto que ya había ocurrido.

Fuentes: NBC News (10 de septiembre de 2026), declaraciones de Joe Benton; cobertura de Storyboard18 y Turkiye Today (septiembre de 2026) sobre las propuestas concretas de divulgación obligatoria.

Quién dijo qué y desde qué posición

Puestas una al lado de la otra, las tres declaraciones describen el mismo problema desde tres ángulos distintos, y esa convergencia es lo que les da peso.

Jacob Coxon, que trabajó en preentrenamiento en OpenAI y en Anthropic, sostuvo que ninguna de las dos empresas actúa de forma responsable, y que están apostando con nuestras vidas.

Joe Benton, que dirigía un equipo de seguridad en Anthropic, sostuvo que el ritmo puede pasar de vertiginoso a incontrolable, y que toda la transparencia actual es voluntaria.

Josh Engels, que trabajaba en seguridad de IA en Google DeepMind, sostuvo que no hay adultos en la sala, y que nadie va a venir a salvarnos.

La segunda declaración es la que más debería pesar en una evaluación profesional, porque proviene de quien dirigía la función de seguridad dentro del laboratorio y tenía acceso directo a los procesos internos de evaluación, de modo que su afirmación sobre el desfase entre capacidad y control constituye un juicio técnico y no una impresión desde afuera.

Fuentes: declaraciones recogidas por NBC News (10 de septiembre de 2026), Associated Press (9 de septiembre de 2026) y la publicación original de Coxon en X.

Lo que diría un lector escéptico, y con razón

Cualquier análisis honesto de este episodio tiene que incorporar las objeciones serias que se le pueden hacer, porque tres renuncias no constituyen por sí solas una prueba de fracaso institucional y conviene decirlo antes de que lo diga alguien más.

La primera objeción apunta al sesgo de selección, porque quien renuncia por motivos de conciencia tiene por definición una visión más pesimista que quien decide quedarse, y de los cientos de personas que trabajan en seguridad dentro de esos laboratorios la enorme mayoría continúa en sus puestos, lo que admite tanto la lectura de que no comparten el diagnóstico como la de que prefieren influir desde adentro.

La segunda objeción es de incentivos y también merece registrarse, ya que Benton y Engels se incorporan a una organización que vive de evaluar a esos mismos laboratorios, de modo que su llamado a fortalecer las evaluaciones independientes coincide con el interés institucional de su nuevo empleador, algo que no invalida el argumento aunque obliga a ponderarlo.

La tercera objeción es la más difícil de responder y opera en sentido contrario, porque si el diagnóstico proviniera solo de críticos externos sería fácil atribuirlo a desconocimiento técnico, y en este caso las tres voces pasaron años dentro de los procesos que ahora describen como insuficientes, de manera que descartar su testimonio exige explicar por qué quienes mejor conocen el sistema serían justamente quienes peor lo entienden.

Fuentes: elaboración propia sobre las objeciones que la cobertura especializada planteó durante la semana; ts2.tech (septiembre de 2026), sobre el contraargumento frente a tratar estas salidas como evidencia de fracaso institucional.

Qué hacemos quienes asesoramos sobre esto en América Latina

Para un abogado, un oficial de cumplimiento o un directivo de la región la pregunta práctica llega enseguida, y consiste en determinar qué cambia en el trabajo diario cuando el equipo de seguridad de un laboratorio afirma que la supervisión no alcanza.

El primer cambio es de fuente, porque una parte importante de la asesoría en gobernanza de inteligencia artificial descansa hoy sobre declaraciones de política publicadas por los propios proveedores, y este episodio recuerda que esas declaraciones carecen de verificación externa obligatoria, de modo que conviene tratarlas como compromisos comerciales y no como garantías técnicas verificadas.

El segundo cambio es contractual y resulta accionable esta semana, ya que si la transparencia del proveedor es voluntaria hacia el regulador, con mayor razón lo es hacia el cliente, y eso vuelve razonable exigir en los contratos obligaciones expresas de notificación de incidentes con plazos, alcance definido y consecuencias, en lugar de confiar en que la información llegará por la vía del comunicado público.

El tercer cambio es de encuadre y afecta a cualquier marco regional que se discuta, porque el pedido de estas tres personas coincide en un punto que las legislaciones latinoamericanas en trámite todavía no resuelven, y consiste en que sin evaluación independiente con acceso real y sin obligación de reportar incidentes y cuasi incidentes, cualquier declaración de principios queda sin mecanismo de verificación.

Fuentes: propuestas de divulgación obligatoria formuladas por Joe Benton a NBC News (10 de septiembre de 2026); análisis propio sobre las implicaciones contractuales y regulatorias para la región.

There are no adults in the room. People are trying their best, but there is no one coming to save us.

Josh Engels, ex investigador de seguridad de IA en Google DeepMind, NBC News, 10 de septiembre de 2026

Oscar Obando Chaves es Chief AI Officer y Legal AI Engineer ecuatoriano, con formación en derecho. Ayuda a las empresas a decidir cómo usan la IA y construye con ella sistemas para el negocio y para lo legal. Chief AI Officer y Senior AI Counsel en HEKA Law, ISO 42001 AI Management Leader e ISO 38507 AI Governance Leader. Publica análisis semanal sobre el impacto real de la IA.