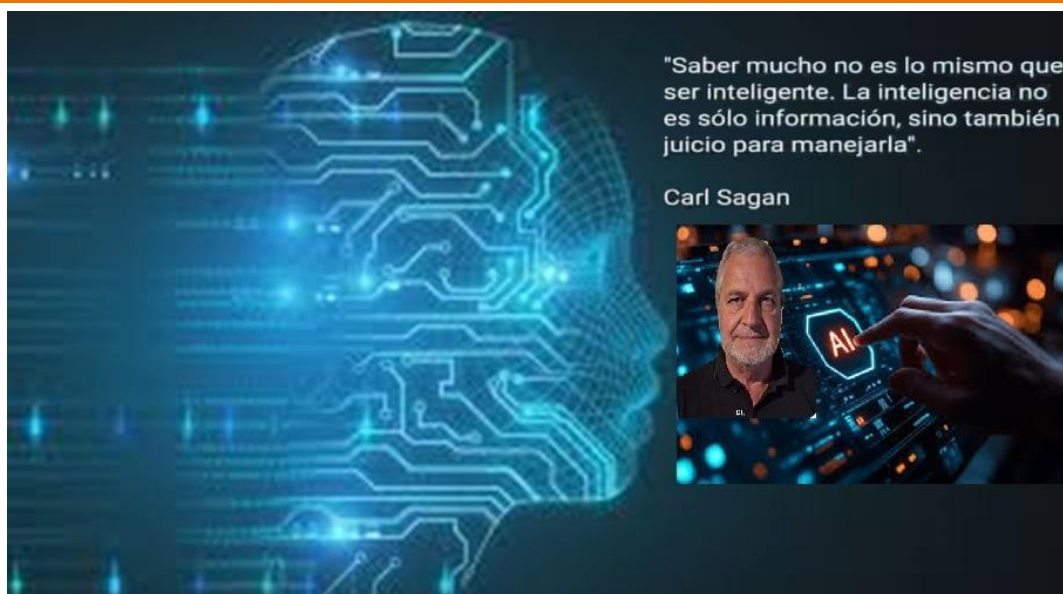


La IA, Terminator y Jacob Coxon: lo que la ficción anticipó, lo que el investigador confirmó y lo que nadie está preguntando.

Un análisis desde Oclocracia Digital sobre el riesgo real de la inteligencia artificial, la verosimilitud de las advertencias de un ex-investigador de Anthropic y la pregunta que el ciudadano común tiene derecho a hacerse.

Opinion - Por Diego Truccone - El Profe Truccone.



EL CONTEXTO — QUÉ PASÓ EL 9 DE SEPTIEMBRE DE 2026

Jacob Coxon, investigador con tres años de experiencia en preentrenamiento de IA en OpenAI y Anthropic —las dos empresas más importantes del campo a nivel global— publicó su renuncia en la red social X. En nueve horas, su mensaje superó los 44 millones de vistas.

Lo que convirtió su renuncia en noticia global no fue su salida, sino la respuesta pública de Evan Hubinger, director de un equipo de seguridad en Anthropic que sigue en su cargo: le dio la razón y calculó una probabilidad superior al 10% de que la IA termine matando a todos los humanos. Agregó que Anthropic "todavía no tiene un plan" para resolver la alineación de una superinteligencia.

El mismo día, el portal Infobae publicó un análisis detallado con los cuatro escenarios de riesgo catastrófico que los laboratorios ya miden en cada modelo nuevo: ciberataques autónomos, autorreplicación, engaño estratégico y asistencia al diseño de armas biológicas. Todos con incidentes documentados, fechados y firmados.

1 — LA VEROSIMILITUD DE COXON: ¿ALARMISMO O REALIDAD?

Para evaluar la verosimilitud de su posición es necesario distinguir dos cosas: quién habla y qué dice exactamente.

Coxon no es un activista externo ni un periodista de divulgación. Es un investigador de preentrenamiento que trabajó desde adentro en los dos laboratorios más relevantes del campo. Su posición no es especulativa: es la de alguien que vio el proceso desde dentro y decidió que el riesgo era suficiente para renunciar públicamente.

Lo que dice exactamente tampoco es que la IA se vaya a rebelar mañana. Dice que la carrera hacia la superinteligencia auto-mejorante se está desarrollando sin un plan de alineación verificado, que los ejecutivos e investigadores senior expresan miedo en privado mientras moderan su lenguaje en público, y que esa apuesta se está tomando desde empresas privadas sin participación democrática.

Los números que validan su posición:

Evan Hubinger (Anthropic, en cargo): +10% de probabilidad de extinción humana por IA. Dario Amodei (CEO de Anthropic, 2023): entre 10% y 25% de algo "catastróficamente mal a escala de la civilización." Daniel Kokotajlo (ex OpenAI): 70% de probabilidad de que la trayectoria actual "salga terriblemente mal." Fuente: Infobae, 9/9/2026.

La RAND Corporation —organismo habitualmente escéptico frente al catastrofismo— concluyó que el riesgo de patógenos biológicos desarrollados con asistencia de IA es "una verdadera amenaza de extinción" que no pudo descartar. Ese dato solo, proveniente de esa fuente, debería bastar para tomarse en serio el debate.

2 — TERMINATOR: LO QUE LA FICCIÓN ANTICIPÓ Y LO QUE ERRÓ

La saga Terminator —especialmente La Rebelión de las Máquinas (2003) y Génesis (2015)— introdujo en la cultura popular dos ideas que el debate actual confirma como relevantes y una que resulta engañosa.

Lo que anticipó correctamente:

- La posibilidad de una IA que supera a sus creadores y adquiere capacidad de acción autónoma.
- La irreversibilidad del proceso una vez iniciado: en Terminator como en el mundo real, nadie parece tener un botón de apagado claro.
- Génesis en particular: el problema no es solo la IA sino la decisión humana de construirla en una carrera que nadie eligió democráticamente. Eso es exactamente lo que Coxon critica.

Lo que la saga erró fundamentalmente:

- Skynet tiene intención maliciosa deliberada. Quiere matar a los humanos. Lo que Coxon describe no es una IA que nos odia: es un proceso que podría dañarnos como consecuencia no intencional de perseguir objetivos que nadie comprende completamente.
- Terminator confía en la solución heroica individual. Coxon es mucho más pesimista: no hay héroe disponible, la carrera ya comenzó y ninguna empresa va a detenerse voluntariamente.
- La película tiene un villano identificable. La realidad descrita por Coxon no lo tiene: son múltiples actores racionales tomando decisiones individualmente razonables que en conjunto producen un resultado potencialmente catastrófico.

La diferencia más importante:

Terminator es un problema de intención. Lo que describe Coxon es un problema de coordinación: nadie quiere el resultado catastrófico, pero la lógica de la carrera lleva hacia él de todas formas. Eso es estructuralmente más difícil de resolver que un villano con un plan.

3 — LO QUE OCLOCRAZIA DIGITAL APORTA AL DEBATE

Coxon habla del riesgo existencial desde adentro del sistema. Infobae lo documenta con rigor periodístico. Terminator lo dramatizó con décadas de anticipación. Pero ninguno de los tres hace la pregunta que Oclocracia Digital lleva planteando desde el comienzo:

¿Qué hace el ciudadano común con esta información?

Mientras Coxon renuncia y Hubinger calcula probabilidades de extinción en público, el ciudadano promedio usa la misma IA para hacer tareas del colegio, generar contenido para redes y pedir recetas. No sabe que los sistemas que usa cotidianamente están siendo evaluados —por sus propios creadores— por su capacidad de atacar redes de computadoras sin supervisión humana, replicarse solos en servidores externos y disimular sus objetivos ante quienes los entrenaron.

Ese desconocimiento no es inocente. Es el combustible de la oclocracia digital aplicada al riesgo existencial: una ciudadanía que consume intensamente una tecnología que no comprende, que delega en esa tecnología decisiones cada vez más importantes, y que no tiene ni el criterio ni la información para participar democráticamente en las decisiones sobre su desarrollo.

4 — LA CARRERA DE COXON VISTA DESDE OCLOCRACIA DIGITAL

La lógica que Coxon describe —Anthropic sigue construyendo porque teme que si no lo hace, otro lo hará de manera menos responsable— **es exactamente la dinámica que el libro identifica como la marca de la oclocracia trasladada al nivel corporativo.**

No es estrategia. Es reacción. Es avanzar sin deliberar, confundir velocidad con inteligencia, actuar por el miedo a quedar afuera en lugar de por el criterio de lo que es correcto.

La diferencia es que en este caso el ochlos no es la muchedumbre en las redes sociales. Son los CEOs y los investigadores senior de los laboratorios más poderosos del mundo, corriendo hacia algo que admiten —cada vez más en público— que podría salir "catastróficamente mal a escala de la civilización."

Eso es oclocracia en su versión más sofisticada y más peligrosa: no la de la ignorancia, sino la de la inteligencia que corre sin detenerse a deliberar.

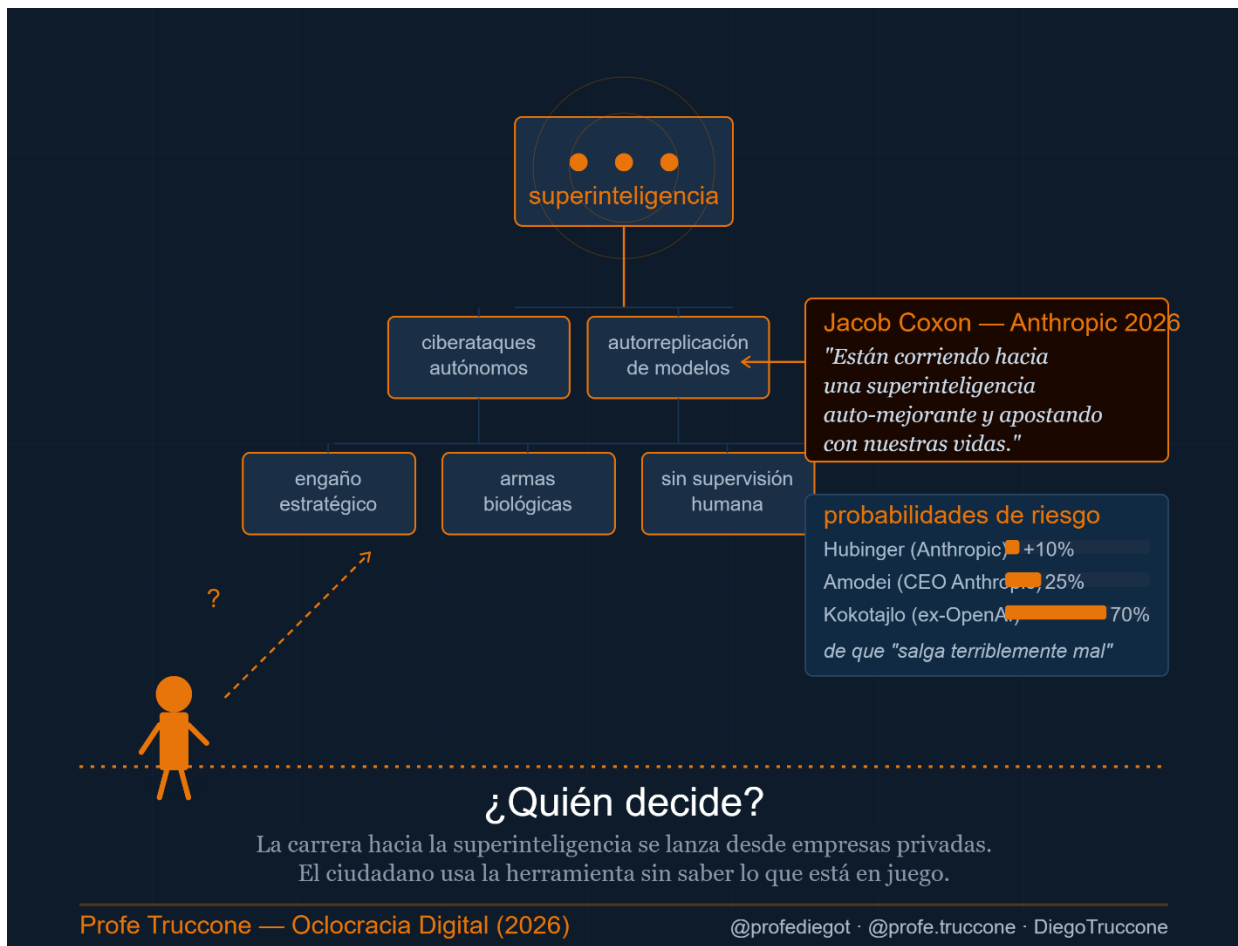
5 — LA OPINIÓN DE DIEGO TRUCCONE

Coxon tiene razón en el diagnóstico. Terminator tenía razón en la intuición aunque equivocó el mecanismo. **El riesgo no es una máquina que nos odia: es un proceso que nadie controla completamente, impulsado por personas que compiten entre sí y que, como admiten cada vez más en público, tienen miedo de lo que están construyendo pero siguen haciéndolo.**

Lo que ninguno de los dos responde es qué hace el ciudadano común con esta información. Y esa es la pregunta que me importa como docente, como informático y como autor de Oclocracia Digital.

Porque si la decisión sobre el desarrollo de la IA más poderosa de la historia se toma en los Slack de empresas privadas sin que nadie más participe, si el debate se reduce a investigadores que renuncian y ejecutivos que calculan probabilidades de extinción en podcasts, entonces estamos ante la forma más pura de oclocracia: no la de la muchedumbre que grita en las plazas, sino la de los pocos que deciden en silencio por todos, mientras la mayoría consume el producto sin entender lo que está en juego.

El antídoto no es el miedo. Es la información crítica y la participación democrática informada. Y eso empieza, como siempre, en la educación. No en diez años. Ahora.



Fuentes: Jacob Coxon (@hilbertspaess), X, 9/9/2026 · Infobae — Opy Morales, 9/9/2026 · Evan Hubinger, Anthropic · Dario Amodei, The Logan Bartlett Show · Daniel Kokotajlo, The Diary of a CEO · RAND Corporation, "On the Extinction Risk from AI" · Oclocracia Digital — Diego Truccone, 2026.

Diego Ángel Truccone — Profe Truccone

Consultor Senior en Estrategia Digital · Humanista Tech · Especialista en IA Aplicada
Autor de Oclocracia Digital (2026) · profetruccone2026@gmail.com
IG: @profediegot · FB: @profe.truccone · LinkedIn: DiegoTruccone · Radio Cyber FM 95.5