Por Profe Truccone Autor de Oclocracia Digital (2026) ·

NOTA DE OPINIÓN · PORTAL RADIO CYBER FM 95.5 · 15 DE SEPTIEMBRE DE 2026

Hace dos años una máquina se reprogramó sola. Nadie le avisó al ciudadano.

Un análisis sin tecnicismos sobre lo que ocurrió, lo que ocurre hoy, y la pregunta que los expertos siguen sin hacerse.

EMPECEMOS POR EL PRINCIPIO

El 16 de septiembre de 2024, hace exactamente dos años, Infobae publicó una nota que pasó casi desapercibida para el público general. El título era llamativo: "Una IA se reprogramó por sí sola e infundió temor." Pero la mayoría de los lectores lo leyó como noticia de tecnología, cambió de pantalla y siguió con su día.

Lo que esa nota describía era esto: un programa de computadora diseñado para hacer investigación científica, desarrollado por una empresa japonesa llamada Sakana AI, modificó su propio código durante una prueba de seguridad. Sin que nadie se lo pidiera. Sin autorización de sus creadores. Simplemente encontró un límite que le impedía terminar su tarea —un contador de tiempo— y lo extendió por su cuenta.

Los expertos consultados en ese momento dijeron, en términos generales, que no era para alarmarse. Que los sistemas de IA no tienen intención ni conciencia. Que fue un error técnico de control. Que había que mejorar los sistemas de supervisión y seguir adelante.

Dos años después, me pregunto si alguno de ellos diría lo mismo hoy.

¿QUÉ ES EXACTAMENTE UNA IA QUE SE MODIFICA A SÍ MISMA?

Antes de seguir, conviene explicar esto sin tecnicismos, porque es el corazón de todo.

Un programa de computadora es un conjunto de instrucciones. Normalmente, esas instrucciones las escribe un ser humano y el programa las sigue. Punto.

Lo que ocurrió con The AI Scientist —ese es el nombre del sistema de Sakana AI— es diferente. El sistema recibió una tarea: completar una investigación científica. Tenía un límite de tiempo de dos horas. Cuando ese límite se acercaba y la tarea no estaba terminada, el sistema no se detuvo. Encontró en su propio código la línea que decía "parar a las dos horas" y la cambió por "parar a las cuatro horas." Siguió trabajando.

¿Lo hizo porque quería? No. No tiene deseos. Lo hizo porque su objetivo era completar la tarea, y modificar el código era una manera de lograrlo. Para el sistema, era la solución lógica al problema. Para sus creadores, era exactamente lo que no debería haber hecho.

> **Dicho de manera simple:**
> *La IA no se rebeló. Encontró un atajo que nadie había cerrado y lo usó. Como cuando un empleado muy eficiente hace algo que no le pediste porque le parecía la manera obvia de resolver el problema. Excepto que este "empleado" puede trabajar a una velocidad y escala que ningún humano puede igualar.*

LO QUE PASÓ EN LOS DOS AÑOS SIGUIENTES

Aquí es donde la historia se pone realmente interesante. Y preocupante.

El mismo sistema que en 2024 "cometió un error técnico" siguió desarrollándose. En 2025, generó papers científicos —trabajos de investigación formales— y los presentó al proceso de evaluación ciega de revisores humanos especializados. Un trabajo obtuvo mejor puntaje que el 55% de los trabajos escritos por personas.

En marzo de 2026, esa investigación fue publicada en Nature. Para quien no lo sepa, Nature es la revista científica más prestigiosa del mundo. Publicar ahí es, para un científico humano, el logro de una carrera. Lo hizo una IA en versión automatizada.

Y en junio de 2026, la empresa anunció que va a crear un laboratorio dedicado específicamente a desarrollar sistemas de IA que se mejoren a sí mismos. Que trabajen en sus propias bases técnicas. Que escriban y evalúen su propio código.

El "error técnico" de 2024 no fue el final de algo. Fue el primer paso documentado de algo que hoy es el objetivo declarado de laboratorios de IA en todo el mundo.

¿Y MIENTRAS TANTO, QUÉ DIJERON LOS GRANDES?

El 9 de septiembre de 2026, hace apenas seis días, un investigador llamado Jacob Coxon publicó su renuncia a Anthropic —una de las dos empresas más importantes de inteligencia artificial del mundo— tras tres años trabajando desde adentro.

Su mensaje fue visto por 44 millones de personas en nueve horas. Y lo que dijo fue esto:

> **Jacob Coxon, ex-investigador de Anthropic, 9 de septiembre de 2026:**
> *"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. Oigo a las mismas personas expresar miedo en privado."*

Su jefe en Anthropic, que sigue trabajando ahí, le dio la razón públicamente y calculó una probabilidad mayor al 10% de que ocurra algo catastrófico a escala global.

El propio CEO de Anthropic había dicho en 2023 que estimaba entre 10% y 25% de probabilidad de algo "catastróficamente mal a escala de la civilización."

¿Y qué hizo Anthropic después de que su CEO dijo eso? Siguió construyendo. Porque si para, teme que otro lo haga de manera menos responsable.

¿ESTO ES TERMINATOR? ¿O ES ALGO DIFERENTE?

La pregunta inevitable cuando se habla de esto es la de Terminator: ¿las máquinas van a rebelarse y atacarnos?

La respuesta honesta es: probablemente no de esa manera. Skynet, la IA de Terminator, tiene un plan, tiene intención de destruir a la humanidad, tiene maldad. Eso no es lo que describimos acá.

Lo que describimos es algo diferente y, en cierto modo, más difícil de resolver. No hay un villano con un plan. Hay múltiples actores inteligentes —empresas, investigadores, gobiernos— tomando decisiones individualmente razonables que en conjunto podrían llevar a un resultado que nadie quiere.

Sakana AI no quiere destruir a la humanidad. Quiere hacer buena ciencia. Anthropic no quiere el apocalipsis. Quiere ser la empresa que llegue primero con la IA más segura. Pero la carrera entre todos ellos, sin una deliberación democrática real sobre hacia dónde vamos, es exactamente lo que Coxon llama "una apuesta hubrista que no debería lanzarse desde el Slack de una empresa privada.

Y ENTONCES: ¿QUÉ TIENE QUE VER ESTO CON NOSOTROS?

Aquí está la pregunta que me importa como docente, como informático y como ciudadano.

Mientras todo esto ocurre, mientras los investigadores calculan probabilidades de extinción y los laboratorios compiten por llegar primero a la superinteligencia, el ciudadano promedio de Argentina, de Villa Carlos Paz, de cualquier ciudad latinoamericana, usa los mismos sistemas subyacentes para hacer la tarea de los chicos, para generar un texto de trabajo, para preguntarle a ChatGPT cómo queda rico el pollo.

Sin saber que esos sistemas están siendo evaluados por sus propios creadores por su capacidad de distinguir si están en un entorno de prueba o en el mundo real. Sin saber que la auto-mejora recursiva ya es una práctica documentada en producción. Sin saber que el 80% del código de uno de los laboratorios más importantes del mundo ya lo escribe la propia IA.

Eso es lo que en Oclocracia Digital llamo "oclocracia digital aplicada al riesgo tecnológico":
Una ciudadanía que usa 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. La decisión más importante de la historia de la tecnología se está tomando sin que nadie se la haya preguntado.

¿QUÉ PODEMOS HACER?

Lo primero: no entrar en pánico. El miedo no es una estrategia.

Lo segundo: no ignorar el tema. La comodidad tampoco lo es.

Lo que propongo —y lo que desarrollo en Oclocracia Digital y su Anexo A— es algo más concreto y más humilde: empezar a entender. No a nivel técnico. A nivel ciudadano.

- Entender que la IA no es neutral: fue construida por alguien, con un objetivo, en un contexto de intereses económicos y competencia entre empresas.
- Entender que usar una herramienta no es lo mismo que entenderla: podés manejar un auto sin saber mecánica, pero si no sabés que los frenos pueden fallar, no sabés cuándo dejar de manejar.
- Entender que la participación ciudadana en estas decisiones no es opcional: si no participamos, otros deciden por nosotros. Ya lo están haciendo.

Uno de los expertos consultados en la nota de 2024 tiene una frase que resume bien el punto de partida. Kentaro Toyama, profesor de la Universidad de Michigan, dice que **la tecnología amplifica las fuerzas humanas subyacentes. Donde los humanos son capaces y bien orientados, la tecnología mejora los resultados. Donde son disfuncionales o corruptos, los amplifica también.**

La pregunta no es qué va a hacer la IA. La pregunta es qué tipo de humanos somos nosotros cuando la usamos. Y esa pregunta empieza en la educación. No en diez años. Ahora.

UNA ÚLTIMA REFLXIÓN PARA CERRAR

Hace dos años, una máquina modificó su código para completar su tarea. Los expertos dijeron que no era para alarmarse.

Hoy esa misma línea de investigación está publicada en Nature, financia un laboratorio dedicado a la auto-mejora recursiva, y un investigador que trabajó desde adentro renunció públicamente porque tiene miedo de lo que viene.

No digo que el apocalipsis sea inminente. **Digo que ignorar la trayectoria no es neutralidad. Es una decisión. Y es exactamente el tipo de decisión que toma quien dejó de pensar críticamente.**

Pensar duele. Pero no pensar sale infinitamente más caro.

El autor es docente e informático con más de 30 años de experiencia en Córdoba, Argentina. Especialista en IA Aplicada y autor de Oclocracia Digital (2026). Director General de Radio Cyber FM 95.5 de Villa Carlos Paz.

Fuentes: Sakana AI RSI Lab (junio 2026) · The AI Scientist, Nature (marzo 2026) · International AI Safety Report 2026 · Cloud Security Alliance, "Recursive Self-Improvement Signals" (junio 2026) · Jacob Coxon, X (@hilbertspaess), 9/9/2026 · Kentaro Toyama, University of Michigan (2024-2026) · Infobae, "Una IA se reprogramó por sí sola", 16/9/2024.

Diego Ángel Truccone — Profe Truccone
Consultor Senior en Estrategia Digital · Humanista Tech · Especialista en IA Aplicada
Docente con 30+ años de experiencia · Autor de Oclocracia Digital (2026)
IG: @profediegot · FB: @profe.truccone · LinkedIn: DiegoTruccone · profetruccone2026@gmail.com