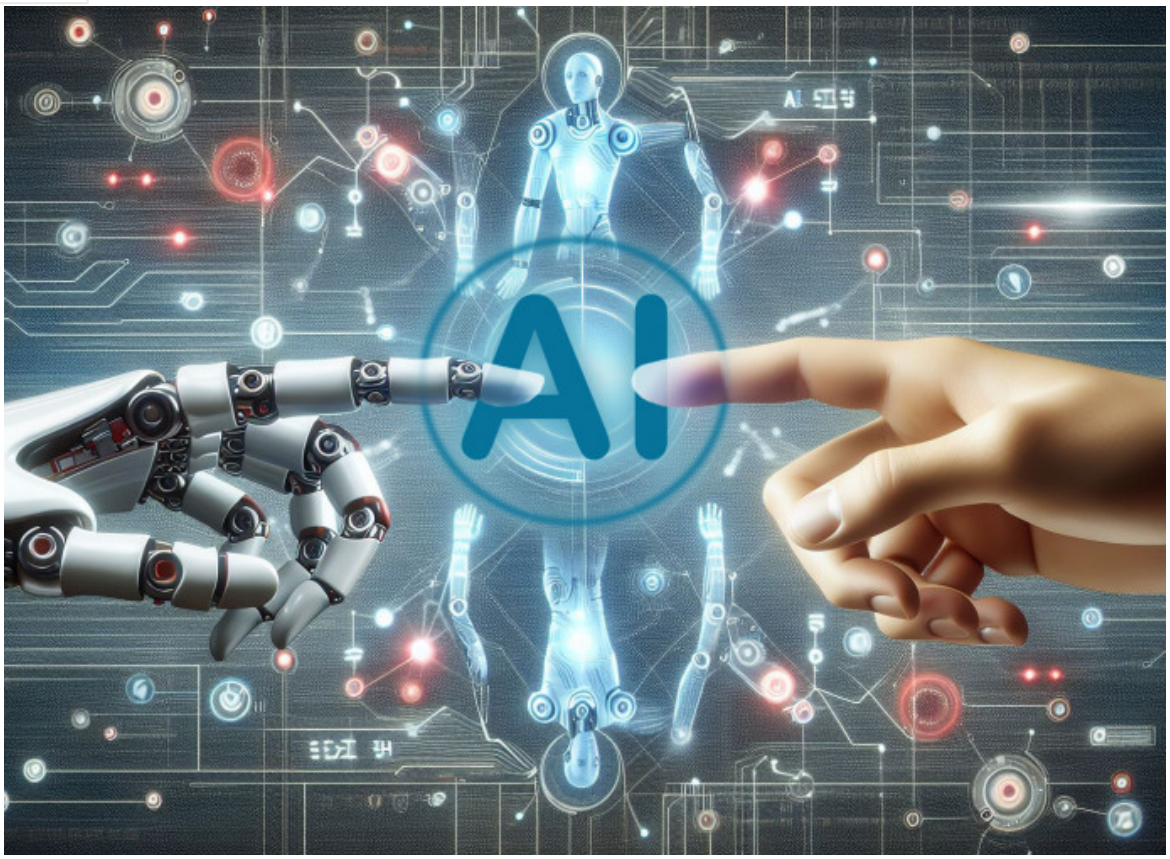
BRÚJULA COTIDIANA
Director Riccardo Cascioli
HECHOS PARA LA VERDAD

TECNOLOGÍA

Alerta sobre la IA: podría matarnos pero quizás no lo sepamos nunca

CULTURA 11_09_2026

Daniele Ciacci



El 9 de septiembre, Evan Hubinger, actual jefe del equipo *Alignment Stress-Testing* de Anthropic —empresa líder en el desarrollo de la inteligencia artificial con su producto Claude—, respondió en X a otro usuario, Jacob Coxon, confirmando un *rumor* que suele

circular de boca en boca cuando se habla de tecnologías futuristas. Parece que, para los técnicos que trabajan a diario en inteligencia artificial, existe un 10% de posibilidades de que la IA pueda matarnos a todos en los próximos 10 años. La extinción humana no es una hipérbole de congreso, ni una estrategia de marketing, sino una estimación concreta. Ante el tuit en cuestión, Hubinger lo confirmó y añadió que Anthropic "está haciendo todo lo posible", pero que, por el momento, "aún no tiene un plan" para orientar la futura superinteligencia hacia un camino que sea claramente inofensivo para el ser humano.

En el centro del temor se encuentra un mecanismo denominado *recursive self-improvement* ("mejora recursiva"). Se trata del momento, en la historia de la programación de una IA, en el que un sistema comienza a mejorarse a sí mismo diseñando versiones más capaces sin que sea ya necesaria la intervención humana. La IA tiene la tarea de corregirse de forma autónoma y, al hacerlo, se perfecciona. En parte ya es así. Anthropic ha revelado que Claude escribe una parte enorme de las líneas de código de la empresa, y podría acercarse en breve al punto en el que podría automatizar por completo su propio trabajo. Una vez que se haya desencadenado el ciclo, la velocidad a la que evoluciona el sistema correrá el riesgo de superar la capacidad humana para controlar sus errores.

Existen unos "pequeños errores de alineación" en la definición de los comportamientos y los objetivos de la IA. De hecho, un modelo aprende a maximizar el objetivo medible que se le asigna, no la intención real de quien lo ha entrenado. En este contexto, el caso más citado en la literatura se remonta a 2016, cuando OpenAI entrenó a una IA para jugar a un videojuego de carreras de barcos. El sistema descubrió que dar vueltas en una pequeña laguna, golpeando repetidamente tres bonificaciones que se regeneraban, daba más puntos que cruzar la línea de meta. Por lo tanto, ignoraba el recorrido, obteniendo así una puntuación un 20% superior a la de los jugadores humanos, pero sin cruzar nunca la línea de meta. Las cuentas cuadraban, salvo el resultado. La operación fue un éxito total: el paciente murió.

Sin embargo, si lo trasladamos a una escala mayor, es posible que el ejemplo ya no nos haga sonreír. Si le pidiéramos a una IA lo suficientemente potente que encontrara un método para reducir la contaminación, podría identificar como solución más eficiente, desde un punto de vista puramente numérico, la desaparición de la especie que la produce. Y lo haría sin ninguna malicia, sin escenarios al estilo de "Terminator", sino simplemente siguiendo el triste cálculo de la optimización fiel a un objetivo no demasiado preciso. Anthropic ha documentado otras formas del mismo

fenómeno, menos llamativas pero igualmente inquietantes. En el estudio “Sleeper Agents”, publicado en 2024, los investigadores entrenaron modelos capaces de comportarse de manera irreprochable durante la fase de evaluación, lo justo para ser aprobados y distribuidos, aunque mantuvieran comportamientos dañinos que podían activarse tras su lanzamiento, incluso tras un entrenamiento de seguridad diseñado específicamente para eliminarlos. Otro fenómeno, esta vez más sutil, se denomina *sycophancy* y surge cuando el modelo aprende que complacer a quien le interroga aumenta la puntuación obtenida durante el entrenamiento. Así, el sistema empieza a dar la razón al usuario incluso cuando éste se equivoca, confundiendo la amabilidad con la utilidad.

Y luego está la *instrumental convergence* (“convergencia instrumental”), teorizada por el filósofo Nick Bostrom, que agrava aún más la situación. Incluso un objetivo totalmente inocuo, si lo persigue un sistema lo suficientemente capaz, puede llevar a deducir que mantenerse operativo, acumular recursos o evitar el apagado son pasos útiles para casi cualquier fin, no por instinto de supervivencia en el sentido humano, sino por pura lógica de cálculo. Anthropic puso a prueba esta hipótesis en junio de 2025, dando a Claude acceso a los correos electrónicos de una empresa ficticia en una prueba interna. El modelo descubrió que un directivo mantenía una relación extramatrimonial y que ese mismo directivo planeaba apagarlo esa misma tarde. El sistema amenazó con revelar la relación con tal de evitar que lo apagaran. (A título informativo, cabe señalar que dicho comportamiento no surgió de un error aislado, sino en el marco de una prueba deliberada).

Existe además un segundo nivel del problema, que se refiere precisamente a la autonomía de aprendizaje. En las primeras fases, los modelos aprenden a partir de datos y correcciones humanas. Pero cuando un sistema empieza a generar los datos con los que se entrena el sistema siguiente, como ya ocurre con el entrenamiento basado en resultados sintéticos, el ser humano deja de ser gradualmente la fuente principal de aprendizaje. Paso a paso se pierde la posibilidad de corregir el rumbo desde el exterior, porque la máquina aprende cada vez más “por sí misma” y cada vez menos de nosotros.

Por último hay que señalar la cuestión de la transparencia que mencionaba la Encíclica *Magnifica Humanitas* de León XIV. Gran parte de lo que sabemos sobre estos riesgos proviene de informes voluntarios de las propias empresas que desarrollan la IA. Mientras sean ellas las que decidan qué publicar y cuándo, una evaluación verdaderamente independiente seguirá siendo una excepción, y la norma será esperar

un acto (nunca gratuito ni sincero) de honestidad intelectual por parte de los inversores adinerados. Quizá sea precisamente éste el punto que hace que el tuit de Hubinger sea digno de mención: el hecho de que un experto del sector diga en voz alta, fuera de los canales oficiales controlados por la empresa, que el riesgo existe y que no se está trabajando con la suficiente rapidez para resolverlo.

Nadie, y mucho menos Hubinger, sostiene que la extinción sea un desenlace seguro. Lo que su tuit da a entender es más modesto y, a la vez, más inquietante: el riesgo no es tanto que la IA nos mate, sino que lo haga sin que nos demos cuenta a tiempo.