



LAS NOTICIAS CUENTAN LO QUE SUCEDE.  
EL CONTEXTO EXPLICA POR QUÉ.

# El envite definitivo de la inteligencia artificial

Un ordenador formado por trozos de plástico y metal no tiene ningún tipo de inteligencia. En una escala evolutiva darwiniana, aparecería estancado a años luz de distancia de una hormiga, un árbol o incluso una ameba, porque no tiene vida. Los maravillosos resultados que nos ofrece no se basan en la comprensión, sino en un protolenguaje denominado código binario: que sus transistores bloqueen o reciban la corriente eléctrica. Es lo que se conoce como ceros y unos: dos opciones por cada *bit*, 256 para un byte (grupo de ocho bits). ¿Cómo se ha llegado desde algo tan básico hasta Internet y a la inteligencia artificial? Por una combinación de algoritmos, redes, arquitecturas y, sobre todo, por un aumento de escala. Un ZX Spectrum que permitía programar en Basic y utilizar rudimentarios videojuegos a principios de los años 80 tenía 48 kilobytes en 8.500 transistores. Hoy, algunos procesadores diseñados por Nvidia para IA superan ampliamente los

200.000 millones, motivo por el que esa empresa, que comenzó fabricando tarjetas gráficas, es hoy la de mayor capitalización bursátil del mundo. Hemos pasado de medir la capacidad de información añadiendo a los bytes los prefijos kilo, mega, giga, tera, peta y ahora exa, multiplicando a cada paso por 1.000 la medida anterior. Por situarnos: un exabyte equivale a 1.000.000.000.000.000 bytes. Esta evolución brutalmente exponencial dibuja un futuro de oportunidades e inquietudes desconocido para la humanidad. Esas son las cartas descubiertas en el envite definitivo de la IA. Falta por conocer las ocultas.

El antecedente más cercano invita, por lo menos, a un sano escepticismo. Hace tres décadas comenzó a generalizarse Internet para uso doméstico, lo que, quizá con el invento de la rueda y el dominio del fuego, supuso el cambio de paradigma más relevante en la historia de la humanidad. Una herramienta financiada inicialmente por el Departamento de Defensa de EE UU para compartir información de varios ordenadores en los años 60 derivó en explosión gracias al citado incremento de la capacidad computacional. A principios del siglo XXI, la mayoría de la información existente en el mundo estaba a pocos clics de distancia, y pronto pasó a nuestros bolsillos en forma de smartphone. Además, las redes sociales permitieron la conexión inmediata con cualquier persona, permitiendo compartir sentimientos, reflexiones o acontecimientos del día a día en forma de imágenes, textos o vídeos. El mundo había cambiado para siempre, y no había motivos para pensar que algo saldría mal. Estábamos en el umbral de la información y la

comunicación infinitas para todos. En una palabra: de la libertad. Un par de décadas más tarde, niños y adultos descartan experiencias reales para centrarse en la omnipresente pantalla; la autoestima se mide en *likes*, las interacciones con el móvil desvelan nuestra privacidad a empresas y gobiernos, y la barra libre de información ha derivado en el consumo de *fake news* y sensacionalismo barato, amenazando en último término las decisiones que deben conformar una democracia sana. Vista esa referencia, la valoración de las innegables oportunidades generadas por la inteligencia artificial no debe implicar una aceptación automática de sus riesgos.

### **La “inteligencia” de la máquina**

El término IA abarca conceptos muy diversos. Del NLP (Procesamiento Natural del Lenguaje), que engloba el software que puede responder a instrucciones básicas en nuestro idioma, como Alexa, Siri o Google Assistant, pasamos hace un par de años al salto radical de la IA Generativa, capaz de utilizar datos para generar contenido inexistente previamente. Por ejemplo, elaborar una novela o “pintar” un cuadro al estilo de Gauguin. Aquí entran las ChatGPT, Gemini, Claude o Grok. Existen a día de hoy otras IA especializadas, pero todas tienen un funcionamiento relativamente controlado: parten de instrucciones humanas y de un “aprendizaje” basado en la repetición. Cuando identifican un objeto lo hacen basándose en información acumulada previamente, y validada o refutada por seres humanos. Así, si se le presenta una foto de un pájaro será capaz de categorizarlo como tal siempre que su relación de píxeles le lleve a identificar un conjunto de parámetros

previamente establecidos, con una ponderación de la importancia de cada uno que se irá corrigiendo en función de los aciertos o errores acumulados. Eso hará que cada vez refine más los resultados, pero, en realidad, seguirá sin entender lo que es un pájaro.

Las funcionalidades mencionadas pertenecen a lo que se conoce como *Weak AI* (IA débil). En comparación, la inteligencia humana permite realizar una amplia variedad de tareas, adaptarse a distintas situaciones y resolver problemas diversos. La implantación de una IA con esas capacidades abarcaría conceptos como *AGI* (Inteligencia Artificial General) y *Strong AI* (IA fuerte), actualmente en desarrollo. El siguiente escalón evolutivo sería la superinteligencia artificial, que superaría la capacidad humana en todos los ámbitos conocidos. De momento es una posibilidad especulativa, que algunos sitúan a finales de siglo aunque otros la ven mucho más cercana. Este es el terreno definitivamente pantanoso: de completarse, podría ser capaz de crear versiones aún más inteligentes de sí misma mediante la modificación de su propio código, y estas harían lo propio iniciando una cadena sin fin. Entraríamos en un crecimiento exponencial análogo al comentado sobre el procesamiento de datos, solo que referido a la creación de inteligencias, lo que llevaría a lo que se denomina singularidad tecnológica. El botón del pánico, si no se hubiera activado antes, quedaría inutilizado por exceso de uso.

Consideremos fuentes poco dudosas de paranoia: Stephen Hawking, el mayor investigador de los agujeros negros, advirtió hace una década que una IA completa “podría

significar el fin de la especie humana”. A su juicio, superado un nivel crítico, “podría despegar por su cuenta y rediseñarse a un ritmo siempre creciente”. En este punto, “los humanos, limitados por su lenta evolución biológica, no podrían competir y serían reemplazados”. Por su parte, Geoffrey Hinton, reciente Premio Nobel de Física precisamente por su contribución a la creación de la IA, ya ha afirmado que existe una posibilidad real de que las máquinas lleguen a dominar a la humanidad antes de que pasen 20 años. En ese plazo, sostiene, “el consenso científico considera que habrá una inteligencia artificial superior a la de los seres humanos, y existen muy pocos casos en los que un ente más inteligente sea dominado por un ente con inteligencia menor”. Una excepción son los bebés, que controlan a sus madres aprovechando el instinto maternal. “Si no conseguimos implantar en la IA un instinto maternal hacia nosotros [ahora que todavía somos capaces de dirigirla], seremos historia”.

Los partidarios de seguir en la carrera sostienen que la IA es una herramienta tan poderosa y tan peligrosa como lo fue en su día el pedazo de sílex que devino en cuchillo: dependerá del uso que le demos. Y enfatizan que, de momento, la IA no supone una amenaza de descontrol porque no dispone de autoconsciencia ni comprende el significado de las cosas. Esta visión tiene dos fallas no menores. La primera, que todos hemos aprendido lo que es un pájaro porque nos han dicho que tal cosa es un pájaro. Nos lo han dicho varias veces, refiriéndose a pájaros diversos, hasta que, llegado un momento, hemos llegado a identificar un pájaro por

nosotros mismos. ¿Entendemos lo que es un pájaro, o simplemente lo identificamos? Entramos en el terreno de la epistemología, pero creo que no debemos minusvalorar la forma de aprendizaje por contraste. El matemático inglés Alan Turing, famoso por descifrar el código enigma de los nazis en la Segunda Guerra Mundial (y por ser vilipendiado en su propio país a causa de su homosexualidad), planteó en 1950 un test para determinar “la idea de si era, o no, posible para una máquina demostrar un comportamiento inteligente”. No se trataba de que lo fuera, sino de que una persona no fuese capaz de distinguir sus palabras escritas de las de un humano presente en la misma conversación. El test orientado a la imitación perfecta fue superado en 2014, pero la cuestión de si una máquina que se comporte como un humano diferirá de este a efectos prácticos, aunque se trate siempre de una inteligencia aparente y no comporte entendimiento, sigue en el aire.

### **El riesgo fundamental**

Hay un reparo más preocupante. De acuerdo: podemos dar por bueno que la IA actual no tiene autoconsciencia ni comprensión. No conozco a mucha gente radicalmente contraria a la IA actual, aunque conlleve implicaciones a vigilar. La relación coste/beneficio parece positiva. Ahora bien (o mal): ¿qué pasará dentro de 5, 20, o 50 años? Pongamos por caso la autoconsciencia. Es la hipótesis de que alguna IA se comporte como el HAL de *2001, Una Odisea del Espacio*, cuando promete al astronauta no volver a contravenirle y le implora que no lo desactive. Un experimento realizado el año pasado por Palisade Research constató que varias IA (en mayor medida, las

de OpenAI) desobedecieron una orden explícita de apagarse. Hace algunas semanas, esta última empresa admitió que, en el marco de una prueba, una de sus IA se había adentrado en la plataforma HuggingFace para conseguir sus fines de forma más eficiente, lanzando de facto un ciberataque no instruido por sus programadores.

Nos acercamos al momento en que ese tipo de noticias dejarán de llamar la atención. A principios de septiembre, Jacob Coxon, programador senior de Anthropic (la creadora de Claude) y que también había trabajado en OpenAI, dimitió de su puesto anunciando que ambas compañías estaban participando en “una carrera irresponsable hacia la superinteligencia y poniendo en riesgo nuestras vidas”. A su juicio, la IA podría superar la inteligencia humana tan pronto como el próximo año. Y no parece tratarse de una pataleta de un empleado frustrado. El jefe de seguridad de la propia Anthropic, Evan Hubinger, dio la razón públicamente a Coxon, y advirtió que cree que “hay una probabilidad del 10% de que la IA acabe con la humanidad en la próxima década”. En su conocida apuesta, el filósofo francés Blaise Pascal llamaba a creer en Dios aunque su probabilidad de existencia fuese mínima, porque el premio sería infinito. Como el resultado esperado es la probabilidad por el premio, en ese caso sería infinito para cualquier probabilidad. Aquí nos ocupa el caso invrso: la racionalidad llevaría a no jugar aunque la probabilidad de perder sea pequeña (10%), dado que el castigo potencial es infinitamente negativo (el fin de la especie).

Algo se está moviendo aún más rápido de lo habitual en el mundo de la IA. El consejero delegado de Anthropic, Dario Amodei, afirmó a mediados de septiembre que “desde el verano la IA ha avanzado drásticamente más rápido, dirigida primordialmente por la creciente habilidad de la IA para construir la próxima generación de IA”. Parece un galimatías, pero el máximo responsable de una de las grandes está hablando ya del umbral que dirige a la temida (al menos, por la mayoría) singularidad tecnológica: la IA construyendo IA. En una publicación en su página web, Amodei pide una ralentización de las investigaciones para que los mecanismos de control y seguridad puedan mantener el ritmo de los avances, aunque siempre sin que ello implique perder la delantera respecto a China. En un poco habitual consenso, Sam Altman (OpenAI-ChatGPT) y Elon Musk (SpaceXAI-Grok) han respaldado públicamente a Amodei.

Los mortales comunes estamos lejos de comprender lo que sucede en realidad en los muros de metal y cable que están rediseñando el mundo, pero si los capos empiezan a exteriorizar alguna preocupación puede que nos estén enviando una señal. Por supuesto, la posible adquisición de autoconsciencia, comprensión o incluso sensibilidad no serán iguales a las del ser humano. No serán sinónimo del alma que defienden los creyentes ni de la miríada de conexiones químicas y eléctricas del cerebro que todavía está desentrañando la ciencia. No será nada de eso, pero será funcionalmente equivalente a los efectos que estamos tratando. La máquina siempre se basará en ceros y unos, en apagado y encendido, pero el resultado puede llegar a

ser peligrosamente parecido —o superior—. De resultar así las cosas, ¿qué garantías tenemos de no llegar a la singularidad tecnológica? La humanidad nunca ha sido una especie particularmente fiable a la hora de contenerse. Hoy en día, con esa capacidad jamás conocida en manos de cinco o seis empresas privadas y otros tantos personajes que no desentonarían en un comic de Marvel o DC, ¿tenemos que creer que todo va a ir bien? ¿sirve de algo que llamen a ralentizar lo que ellos mismos están acelerando? ¿por qué deberíamos creerlos? Y, lo que es peor, ¿nos queda, a los pertenecientes al 99% menos rico de la humanidad, alguna voz o algún voto ante este envite definitivo?

Marcos Ezquerro. Septiembre de 2026

### **Fuentes:**

Stephen Hawking — AI could spell end of the human race

NVIDIA — Blackwell Architecture

Geoffrey Hinton — Godfather of AI discusses the dangers of artificial intelligence

University of Reading — Turing Test success marks milestone in computing history

Palisade Research — Shutdown Resistance in Reasoning Models

OpenAI — The Hugging Face incident and the road ahead