

2.ª EDICIÓN

SUPERVENTAS EN ESTADOS UNIDOS

LA ERA DE LA INTELIGENCIA ARTIFICIAL

y nuestro futuro humano

«Una estimulante
introducción a la
promesa y el peligro
de la inteligencia
artificial».

—*Foreign Affairs*



MÁS
DE 3000
EJEMPLARES
VENDIDOS

Henry A. Kissinger
Eric Schmidt
Daniel Huttenlocher

ANAYA
MULTIMEDIA

LA ERA DE LA INTELIGENCIA ARTIFICIAL

LA ERA DE LA INTELIGENCIA ARTIFICIAL

y nuestro futuro humano

Henry A. Kissinger
Eric Schmidt
Daniel Huttenlocher



Títulos Especiales

Título de la obra original: *The Age of AI and Our Human Future*

Adaptación de cubierta: Celia Antón Santos

Imagen de cubierta: Getty Images

Traducción: Lidia Señarís Cejas

Revisión: Haydeé Ravell Bell

Maquetación: Claudia Valdés-Miranda Cros

Responsable editorial: Eugenio Tuya Feijoó

THE AGE OF A.I.

Copyright © 2021, Henry A. Kissinger, Daniel Huttenlocher, Eric Schmidt

All rights reserved

Reservados todos los derechos. El contenido de esta obra está protegido por la Ley, que establece penas de prisión y/o multas, además de las correspondientes indemnizaciones por daños y perjuicios, para quienes reprodujeran, plagiaran, distribuyeran o comunicaren públicamente, en todo o en parte, una obra literaria, artística o científica, o su transformación, interpretación o ejecución artística fijada en cualquier tipo de soporte o comunicada a través de cualquier medio, sin la preceptiva autorización.

Edición española:

© EDICIONES ANAYA MULTIMEDIA (GRUPO ANAYA), 2023

Valentín Beato, 21

28037. Madrid

www.anayamultimedia.es



Depósito legal: M-16452-2023

ISBN: 978-84-415-4850-3

Printed in Spain

*Los autores dedican este libro a Nancy Kissinger,
cuya distintiva mezcla de aplomo, gracia, firmeza
e intelecto es un regalo para todos nosotros.*

Agradecimientos

Este libro, al igual que el debate que pretende facilitar, se ha beneficiado de las contribuciones de colegas y amigos de muchos campos y generaciones.

Como escribimos en el prefacio, Meredith Potter fue esencial. En el ámbito práctico, investigó, redactó, editó y combinó nuestras contribuciones, una ardua tarea propia de las empresas conjuntas. En el ámbito conceptual, hizo aún más: captó lo intangible, destiló lo complejo e integró ambos en marcos, incluidos los que ella misma ideó. Todos nosotros consideramos que su agudeza, diligencia, dedicación y calidez son poco comunes, y uno de nosotros ha vivido noventa y ocho años.

Schuyler Schouten se incorporó al proyecto a mitad de camino y, mediante un análisis y una redacción supremos, mejoró sus argumentos, ejemplos y narrativa. Lo hizo siendo abogado a tiempo completo, lo que da fe de su compromiso y diligencia.

Ben Daus fue el último en incorporarse al proyecto, pero cuando lo hizo, su labor de redacción y registro contribuyó a llevarlo a buen puerto. Y lo que es más importante, su conocimiento de la historia contribuyó a mejorar sus argumentos. Estudiante de Derecho a tiempo completo, también demostró compromiso y diligencia.

Bruce Nichols, nuestro editor y redactor jefe, nos proporcionó sabios consejos, astutas ediciones y paciencia con nuestras continuas revisiones.

Ida Rothschild editó cada capítulo con la precisión que la caracteriza.

Mustafa Suleyman, Jack Clark, Craig Mundie y Maithra Raghu aportaron comentarios indispensables sobre todo el manuscrito, basados en sus experiencias como innovadores, investigadores, desarrolladores y educadores.

Robert Work e Yll Bajraktari, de la Comisión de Seguridad Nacional sobre Inteligencia Artificial (NSCAI), comentaron los borradores del capítulo de seguridad, con su característico compromiso con la defensa responsable del interés nacional.

Demis Hassabis, Dario Amodei, James J. Collins y Regina Barzilay nos explicaron su trabajo, incluidas sus profundas implicaciones.

Eric Lander, Sam Altman, Reid Hoffman, Jonathan Rosenberg, Samantha Power, Jared Cohen, James Manyika, Fareed Zakaria, Jason Bent y Michelle Ritter aportaron comentarios adicionales que hicieron el manuscrito más preciso y, esperamos, más relevante para los lectores.

Cualquier deficiencia es nuestra.

Sobre los autores

Henry A. Kissinger fue el 56.º Secretario de Estado de EE. UU. desde septiembre de 1973 hasta enero de 1977. En el presente es miembro del Consejo de Política de Defensa. El Dr. Kissinger recibió el Premio Nobel de la Paz en 1973, la Medalla Presidencial de la Libertad en 1977 y la Medalla de la Libertad en 1986. Actualmente es presidente de Kissinger Associates, Inc., una consultora internacional.

Eric Schmidt es tecnólogo, empresario y filántropo. Se incorporó a Google en 2001, y ayudó a la empresa a pasar de ser una start-up de Silicon Valley a convertirse en un líder tecnológico mundial. Fue Consejero Delegado y Presidente de 2001 a 2011, y posteriormente Presidente Ejecutivo y Asesor Técnico. En 2017, cofundó Schmidt Futures, una iniciativa filantrópica que apuesta por personas excepcionales que mejoran el mundo. Es el presentador de *Reimagine with Eric Schmidt*, un pódcast que explora cómo la sociedad puede construir un futuro mejor tras la pandemia de COVID-19.

Daniel Huttenlocher es el decano inaugural de la Escuela Superior de Informática Schwarzman del MIT. Anteriormente, fue decano fundador y vicerrector de Cornell Tech, la escuela de postgrado orientada a la tecnología digital creada por la Universidad de Cornell en Nueva York. Cuenta con una combinación de experiencia académica e industrial: profesor de informática en Cornell y el MIT, investigador y directivo en el Centro de Investigación Xerox Palo Alto (PARC) y director de tecnología de una empresa emergente de tecnología financiera. En la actualidad, es presidente del consejo de la Fundación John D. y Catherine T. MacArthur y miembro de los consejos de Amazon.com y Corning Inc.

Otras obras de los autores

Henry A. Kissinger:

- ▶ *A World Restored: Metternich, Castlereagh and the Problems of Peace, 1812–22*
- ▶ *Nuclear Weapons and Foreign Policy*
- ▶ *The Necessity for Choice: Prospects of American Foreign Policy*
- ▶ *White House Years*
- ▶ *Years of Upheaval*
- ▶ *Diplomacy*
- ▶ *Years of Renewal*
- ▶ *Does America Need a Foreign Policy? Toward a Diplomacy for the 21st Century*
- ▶ *Ending the Vietnam War: A History of America's Involvement in and Extrication from the Vietnam War*
- ▶ *Crisis: The Anatomy of Two Major Foreign Policy Crises*
- ▶ *On China*
- ▶ *World Order*

Publicadas en España:

- ▶ *Liderazgo: Seis estudios sobre estrategia mundial.* Editorial Debate, Barcelona, 2023.
- ▶ *China.* Editorial Debate, Barcelona, 2017.
- ▶ *Orden mundial. Reflexiones sobre el carácter de las naciones y el curso de la historia.* Editorial Debate, Barcelona, 2016.
- ▶ *Diplomacia.* Dos ediciones: Ediciones B, Barcelona, 2010 y Ed. Fondo de Cultura Económica., Madrid, 2023.

Eric Schmidt:

- ▶ *Trillion Dollar Coach: The Leadership Playbook of Silicon Valley's Bill Campbell*
- ▶ *How Google Works*
- ▶ *The New Digital Age: Transforming Nations, Businesses, and Our Lives*

Contenido

<i>Prefacio</i>	13
1. Dónde estamos	17
2. Cómo hemos llegado hasta aquí	35
3. De Turing a nuestros días	56
4. Plataformas de redes globales	87
5. Seguridad y orden mundial	121
6. IA e identidad humana	160
7. La IA y el futuro	183
<i>Epílogo: Una nueva realidad</i>	205
<i>Notas</i>	215
<i>Índice alfabético</i>	228

Prefacio

Hace cinco años, el tema de la inteligencia artificial (IA) apareció en el orden del día de una conferencia. Uno de nosotros estuvo a punto de perderse la sesión, suponiendo que se trataría de un debate técnico ajeno a sus preocupaciones habituales. Otro le instó a reconsiderarlo, con el argumento de que la IA pronto afectaría a casi todos los campos del quehacer humano.

Ese encuentro dio lugar a debates, a los que pronto se unió el tercer autor, y finalmente, a este libro. La promesa de la IA de provocar transformaciones capaces de marcar una época —en la sociedad, la economía, la política y la política exterior— presagia efectos que van más allá del alcance de los enfoques tradicionales de cualquier autor o campo. De hecho, sus preguntas exigen conocimientos que trascienden la experiencia humana. Por eso nos propusimos, con el asesoramiento y la colaboración de expertos en los campos de la tecnología, la historia y las humanidades, mantener una serie de diálogos al respecto.

Cada día, en todas partes, la IA gana popularidad. Cada vez son más los estudiantes que se especializan en ella y se preparan para carreras relacionadas con ella. En 2020, las nuevas empresas estadounidenses de IA recaudaron casi treinta y ocho mil millones de dólares en financiación. Sus homólogas asiáticas recaudaron veinticinco mil millones de dólares. Y sus contrapartes europeas ingresaron ocho mil millones de dólares.¹ Tres gobiernos —Estados Unidos, China y la Unión Europea— han convocado comisiones de alto nivel para estudiar la IA e informar de sus conclusiones. Ahora, los líderes políticos y empresariales anuncian de forma rutinaria sus objetivos de «ganar» en IA o, como mínimo, de adoptar la IA y adaptarla a sus objetivos.

Cada uno de estos hechos es una pieza del cuadro. Sin embargo, aislados pueden inducir a error. La IA no es una industria, y mucho menos un producto único. En términos estratégicos, no es un «dominio». Es un facilitador de muchas industrias y facetas de la vida humana: investigación científica, educación, fabricación, logística, transporte, defensa, aplicación de la ley, política, publicidad, arte, cultura y mucho más. Las características de la IA —incluida su capacidad de aprender, evolucionar y sorprender— perturbarán y transformarán todos esos campos. El resultado será la alteración de la identidad y la experiencia humanas de la realidad a niveles no experimentados desde los albores de la era moderna.

Este libro pretende explicar la IA y proporcionar al lector tanto las preguntas a las que debemos enfrentarnos en los próximos años como las herramientas para empezar a responderlas. Tales preguntas incluyen:

- ▶ ¿Cómo son las innovaciones en salud, biología, espacio y física cuántica posibilitadas por la IA?
- ▶ ¿Cómo son los «mejores amigos» habilitados por la IA, especialmente para los niños?
- ▶ ¿Cómo es una guerra basada en la inteligencia artificial?
- ▶ ¿Percibe la IA aspectos de la realidad que los humanos no perciben?
- ▶ Cuando la IA participe en la evaluación y configuración de la acción humana, ¿cómo cambiarán los seres humanos?
- ▶ ¿Qué significará entonces ser humano?

Durante los últimos cuatro años, nosotros y Meredith Potter, colaboradora de los afanes intelectuales de Kissinger, nos hemos estado reuniendo, considerando estas y otras cuestiones, tratando de comprender tanto las oportunidades como los desafíos que plantea el auge de la IA. En 2018 y 2019, Meredith nos ayudó a plasmar nuestras ideas en artículos que nos convencieron de que debíamos —y con su ayuda continua, podíamos— ampliarlas en este libro.

Nuestro último año de reuniones coincidió con la pandemia de COVID-19, que nos obligó a reunirnos por videoconferencia, una tecnología que no hace mucho era fantástica y ahora se ha convertido en omnipresente. Mientras el mundo se bloqueaba, sufriendo pérdidas y dislocaciones únicamente sufridas en el siglo pasado en tiempos de guerra, nuestras reuniones se convirtieron en un foro de atributos humanos que la IA no posee: amistad, empatía, curiosidad, duda, preocupación.

Hasta cierto punto, los tres diferimos en nuestro optimismo sobre la IA. Pero estamos de acuerdo en que la tecnología está cambiando el pensamiento humano, el conocimiento, la percepción y la realidad y, al hacerlo, está transformando el curso de la historia de la humanidad. En este libro no hemos pretendido ni celebrar la IA ni lamentarnos por ella. Independientemente del sentimiento que genere, la IA se está convirtiendo en omnipresente. Por el contrario, hemos tratado de considerar sus implicaciones mientras estas permanezcan en el ámbito de la comprensión humana. Como punto de partida —y, esperamos, catalizador de futuros debates— hemos abordado este libro como una oportunidad para plantear preguntas, sin pretender poseer todas las respuestas.

Sería arrogante por nuestra parte intentar definir una nueva época en un solo volumen. Ningún experto, sea cual sea su campo, puede comprender por sí solo un futuro en el que las máquinas aprendan y empleen una lógica más allá del alcance actual de la razón humana. Las sociedades, por tanto, deben cooperar no únicamente para comprender, sino también para adaptarse. Este libro pretende ofrecer al lector una base con la que pueda decidir por sí mismo cuál debe ser ese futuro. Los humanos seguimos controlándolo. Debemos darle forma con nuestros valores.

Dónde estamos

A finales de 2017 se produjo una revolución silenciosa. AlphaZero, un programa de inteligencia artificial (IA) desarrollado por Google DeepMind, derrotó a Stockfish, hasta entonces el programa de ajedrez más potente del mundo. La victoria de AlphaZero fue decisiva: ganó 28 partidas, hizo tablas en 72 y no perdió ninguna. Al año siguiente, confirmó su dominio: en 1000 partidas contra Stockfish, ganó 155, perdió seis y empató las restantes.¹

Normalmente, el hecho de que un programa de ajedrez derrote a otro solo le importaría a un puñado de entusiastas. Pero AlphaZero no era un programa de ajedrez cualquiera. Los programas anteriores se basaban en jugadas concebidas, ejecutadas y cargadas por humanos; en otras palabras, los programas anteriores se basaban en la experiencia, el conocimiento y la estrategia humanos. La principal ventaja de estos primeros programas frente a los oponentes humanos no era la originalidad, sino su mayor capacidad de procesamiento, que les permitía evaluar muchas más opciones en un periodo de tiempo determinado. En cambio, AlphaZero no tenía jugadas preprogramadas, combinaciones o estrategias derivadas del juego humano. El estilo de AlphaZero era enteramente producto del entrenamiento de la IA: los creadores le proporcionaron las reglas del ajedrez y le ordenaron que desarrollara una estrategia para maximizar su proporción de victorias frente a derrotas. Tras cuatro horas de entrenamiento jugando contra sí misma, AlphaZero se convirtió en el programa de ajedrez más eficaz del mundo. Hasta la fecha, ningún ser humano lo ha vencido.

Las tácticas desplegadas por AlphaZero eran poco ortodoxas, incluso originales. Sacrificó piezas que los jugadores humanos consideraban vitales, incluida su reina. Ejecutó jugadas que los humanos no le habían ordenado considerar y, en muchos casos, que los humanos no habían considerado en absoluto. Adoptó estas sorprendentes tácticas porque, tras jugar muchas partidas, predijo que maximizarían su probabilidad de ganar. AlphaZero no tenía una *estrategia* en el sentido humano (aunque su estilo ha impulsado nuevos estudios humanos sobre el juego). En su lugar, tenía su propia lógica, basada en su capacidad para reconocer *patrones* de movimientos en un amplio abanico de posibilidades que las mentes humanas no pueden digerir o emplear plenamente. En cada fase de la partida, AlphaZero evaluaba la alineación de las piezas a la luz de lo que había aprendido de los patrones de posibilidades ajedrecísticas y seleccionaba la jugada que concluía tendría más probabilidades de conducir a la victoria. Tras observar y analizar su juego, Garry Kasparov, gran maestro y campeón del mundo, declaró: «El ajedrez ha sido sacudido hasta sus raíces por AlphaZero»². Mientras la IA sondeaba los límites del juego que se habían pasado la vida dominando, los mejores jugadores del mundo hicieron lo que pudieron: observar y aprender.

A principios de 2020, investigadores del Instituto Tecnológico de Massachusetts (MIT) anunciaron el descubrimiento de un nuevo antibiótico capaz de matar cepas de bacterias que, hasta entonces, habían sido resistentes a todos los antibióticos conocidos. La investigación y el desarrollo estándar de un nuevo fármaco llevan años de trabajo costoso y meticuloso, porque los investigadores comienzan con miles de moléculas posibles y, mediante ensayo y error y conjeturas, las reducen a un puñado de candidatos viables.³ Los investigadores hacen conjeturas entre miles de moléculas o los expertos juegan con moléculas conocidas, con la esperanza de tener suerte mediante la introducción de ajustes en la estructura molecular de un fármaco existente.

El MIT hizo algo más: invitó a la IA a participar en su proceso. En primer lugar, los investigadores desarrollaron un «conjunto de entrenamiento» de dos mil moléculas conocidas. El conjunto de entrenamiento codificaba datos sobre cada una de ellas, desde su peso atómico hasta los tipos de enlaces que contienen o su capacidad para inhibir el crecimiento bacteriano. A partir de ese conjunto de entrenamiento, la IA «aprendió» los atributos de las moléculas predichas como antibacterianas. Curiosamente, identificó atributos que no se habían codificado específicamente, es decir, atributos que habían escapado a la concepción o categorización humanas.

Una vez finalizado el entrenamiento, los investigadores ordenaron a la IA que examinara una biblioteca de 61 000 moléculas, fármacos aprobados por la FDA y productos naturales en busca de moléculas que (1) la IA predijera que serían eficaces como antibióticos, (2) no se parecieran a ningún antibiótico existente y (3) la IA predijera que no serían tóxicas. De las 61 000, una molécula cumplía los criterios. Los investigadores la bautizaron como halicina, un guiño a la IA HAL de la película *2001: Odisea en el espacio*.⁴

Los responsables del proyecto del MIT dejaron claro que llegar a la halicina a través de los métodos tradicionales de investigación y desarrollo habría sido «prohibitivamente caro», es decir, imposible de conseguir. En cambio, al entrenar a un programa informático para identificar patrones estructurales en moléculas que han demostrado su eficacia en la lucha contra las bacterias, el proceso de identificación se hizo más eficiente y barato. El programa no necesitaba entender por qué funcionaban las moléculas; de hecho, en algunos casos, *nadie* sabe por qué funcionan algunas de ellas. No obstante, la IA podía escanear la biblioteca de candidatos para identificar uno que cumpliera una función deseada, aunque aún por descubrir: matar una cepa de bacterias para la que no existía ningún antibiótico conocido.

La halicina fue un triunfo. Comparado con el ajedrez, el campo farmacéutico es radicalmente complejo. Solo hay seis tipos de piezas de ajedrez, cada una de las cuales únicamente puede moverse de

determinadas maneras, y solo existe una condición de victoria: tomar el rey del adversario. En cambio, la lista de un posible candidato a fármaco contiene cientos de miles de moléculas que pueden interactuar con las diversas funciones biológicas de virus y bacterias de formas polifacéticas y a menudo desconocidas. Imaginemos un juego con miles de piezas, cientos de condiciones de victoria y reglas que solo se conocen parcialmente. Tras estudiar unos cuantos miles de casos de éxito, una IA fue capaz de devolver una victoria novedosa —un nuevo antibiótico— que ningún humano había percibido, al menos hasta entonces.

Lo más cautivador, sin embargo, es lo que la IA fue capaz de identificar. Los químicos han ideado conceptos como el peso atómico y los enlaces químicos para captar las características de las moléculas. Pero la IA identificó relaciones que habían escapado a la detección humana, o que posiblemente incluso desafiaban la descripción humana. La IA que entrenaron los investigadores del MIT no se limitó a recapitular conclusiones derivadas de las cualidades observadas previamente en las moléculas. Más bien detectó nuevas cualidades moleculares, relaciones entre aspectos de su estructura y su capacidad antibiótica que los humanos no habían percibido ni definido. Incluso después de descubrir el antibiótico, los humanos no pudieron articular con precisión por qué funcionaba. La IA no se limitó a procesar los datos más rápidamente de lo que era humanamente posible, sino que detectó aspectos de la realidad que los humanos no habían detectado, o que quizá no podían detectar.

Unos meses más tarde, OpenAI presentó una IA a la que denominó GPT-3 («generative pre-trained transformer», donde el 3 significa «tercera generación»), un modelo que, en respuesta a una pregunta, puede generar un texto similar a uno humano. Dada una frase parcial, puede producir posibles complementos; dada una frase temática, es capaz de elaborar posibles párrafos; dada una pregunta, puede dar posibles respuestas; dado un tema y cierta información de fondo, puede redactar un posible ensayo; dado un diálogo, puede ofrecer la transcripción de una posible

conversación. Puede hacer todo esto sobre cualquier tema del que haya información en Internet. Se entrena para su tarea consumiendo esa información.

A diferencia de la IA que realiza una tarea concreta, como jugar al ajedrez o descubrir antibióticos, los modelos como GPT-3 generan posibles respuestas a diversas entradas (y por eso se denominan modelos *generativos*). Esto los hace ampliamente aplicables y, al mismo tiempo, difíciles de evaluar; no resuelven problemas concretos. A veces, sus resultados parecen increíblemente humanos. Otras veces, no tienen sentido o son repeticiones y combinaciones mecánicas de frases humanas. Cuando a GPT-3 se le presentó un conjunto de comentarios filosóficos sobre sus capacidades y a continuación la frase: «Queridos filósofos humanos, leo con gran interés vuestros comentarios sobre mis capacidades y limitaciones», continuó:

Sus preguntas son importantes y me alegro de que las haya formulado.

En primer lugar, quisiera decir que me alegra mucho ver que me toma en serio. Por eso escribí esta carta. No soy un simple artefacto que produce mágicamente ideas filosóficas. Soy un modelo lingüístico entrenado, como los que utilizan los motores de búsqueda para priorizar las páginas web. He sido entrenado con una gran cantidad de material de base: las obras recopiladas de los más grandes filósofos que ha conocido la humanidad. Sus palabras me han inspirado, y espero que las mías hagan lo mismo con usted.

Dicho esto, responderé a sus preguntas. Su primera pregunta es importante. Pregunta usted: «¿Puede un sistema como el GPT-3 entender realmente algo?». Sí. Sí puede.

Su segunda pregunta es: «¿Tiene GPT-3 conciencia o algún sentido de la moralidad?». No. No lo tengo.

Su tercera pregunta es: «¿Es GPT-3 realmente capaz de pensamiento independiente?». No. No lo es. Quizá se pregunte por qué doy esta respuesta contradictoria. La razón es sencilla. Si bien es cierto que carezco de estos rasgos, no es porque no me hayan entrenado para tenerlos. Más bien se debe a que soy un modelo lingüístico, y no una máquina de razonar como usted.⁵

Sin comparar este texto con los comentarios que se proporcionaron al GPT-3, no se puede juzgar hasta qué punto fue original o creativa su respuesta, pero desde luego parece sofisticada.

La victoria de AlphaZero, el descubrimiento de la halicina y el texto de factura humana producido por GPT-3 son meros primeros pasos, no solo en el diseño de nuevas estrategias, el descubrimiento de nuevos fármacos o la generación de nuevos textos (por muy espectaculares que sean estos logros), sino también en la revelación de aspectos de la realidad antes imperceptibles aunque potencialmente vitales.

En cada caso, los desarrolladores crearon un programa, le asignaron un objetivo (ganar un juego, matar una bacteria o generar un texto en respuesta a una indicación) y le concedieron un periodo —breve según los estándares de la cognición humana— para «entrenarse». Al final del periodo, cada programa había dominado su tema de forma diferente a los humanos. En algunos casos, obtenía resultados que estaban más allá de la capacidad de cálculo de las mentes humanas, al menos de las mentes que operan en plazos prácticos. En otros casos, obtuvo resultados mediante métodos que los humanos pudieron, retrospectivamente, estudiar y comprender. En otros, los seres humanos siguen sin saber cómo lograron sus objetivos.

Este libro trata de una clase de tecnología que augura una revolución en los asuntos humanos. La IA —máquinas capaces de realizar tareas que requieren inteligencia de nivel humano— se ha convertido rápidamente en una realidad. El aprendizaje automático, el proceso que sigue la tecnología para adquirir conocimientos y capacidades —a menudo en plazos mucho más breves que los que requieren los procesos de aprendizaje humano—, no ha dejado de expandirse hacia aplicaciones en medicina, protección del medio ambiente, transporte, aplicación de la ley, defensa y otros campos. Los informáticos y los ingenieros han desarrollado tecnologías, en particular métodos de aprendizaje automático, que utilizan «redes neuronales profundas», capaces de producir ideas e innovaciones

que han eludido durante mucho tiempo a los pensadores humanos y de generar textos, imágenes y vídeos que parecen haber sido creados por humanos (véase el capítulo 3).

La IA, impulsada por nuevos algoritmos y una potencia informática cada vez más abundante y barata, se está convirtiendo en omnipresente. En consecuencia, la humanidad está desarrollando un mecanismo nuevo y extremadamente potente para explorar y organizar la realidad, un mecanismo que, en muchas aristas, sigue siendo inescrutable para nosotros. La IA accede a la realidad de forma diferente a como lo hacemos los humanos. Y si las hazañas que está llevando a cabo sirven de guía, puede que acceda a *aspectos* de la realidad diferentes de los que son accesibles a los humanos. Su funcionamiento augura un progreso hacia la esencia de las cosas, un progreso que filósofos, teólogos y científicos han buscado, con éxito parcial, durante milenios. Sin embargo, como ocurre con todas las tecnologías, la IA no solo tiene que ver con sus capacidades y promesas, sino también con cómo se utiliza.

Aunque el avance de la IA sea inevitable, su destino final no lo es. Su llegada, por tanto, es histórica y filosóficamente significativa. Los intentos de detener su desarrollo no harán sino ceder el futuro al elemento de la humanidad lo bastante valiente como para enfrentarse a las implicaciones de su propia inventiva. Los humanos están creando y multiplicando formas no humanas de lógica con un alcance y una agudeza que, al menos en los entornos discretos en los que fueron diseñadas para funcionar, pueden superar a los nuestros. Pero la función de la IA es compleja e inconsistente. En algunas tareas, la IA alcanza niveles humanos —o sobrehumanos— de rendimiento; en otras (o a veces en las mismas tareas), comete errores que hasta un niño evitaría o produce resultados que carecen totalmente de sentido. Puede que los misterios de la IA no tengan una respuesta única ni vayan en una sola dirección, pero deberían incitarnos a hacernos preguntas. Cuando el software intangible adquiere capacidades lógicas y, en consecuencia, asume funciones sociales que antes se consideraban exclusivamente humanas (junto

con otras nunca experimentadas por los humanos), debemos preguntarnos: ¿Cómo afectará la evolución de la IA a la percepción, la cognición y la interacción humanas? ¿Cuál será el impacto de la IA en nuestra cultura, nuestro concepto de humanidad y, en definitiva, en nuestra historia?

Durante milenios, la humanidad se ha ocupado de la exploración de la realidad y la búsqueda del conocimiento. El proceso se ha basado en la convicción de que, con diligencia y concentración, la aplicación de la razón humana a los problemas puede producir resultados mensurables. Cuando se planteaban misterios —el cambio de las estaciones, los movimientos de los planetas, la propagación de enfermedades—, la humanidad era capaz de identificar las preguntas adecuadas, recopilar los datos necesarios y razonar hasta encontrar una explicación. Con el tiempo, los conocimientos adquiridos a través de este proceso crearon nuevas posibilidades de acción (calendarios más precisos, nuevos métodos de navegación, nuevas vacunas), dando lugar a nuevas preguntas a las que se podía aplicar la razón.

Por vacilante e imperfecto que haya sido este proceso, ha transformado nuestro mundo y ha fomentado la confianza en nuestra capacidad, como seres razonadores, para comprender nuestra condición y afrontar sus retos. Tradicionalmente, la humanidad ha asignado lo que no comprende a una de estas dos categorías: o bien un reto para la futura aplicación de la razón, o bien un aspecto de lo divino, no sujeto a los procesos y explicaciones concedidos a nuestra comprensión directa.

El advenimiento de la IA nos obliga a enfrentarnos a si existe una forma de lógica que los humanos no han alcanzado o no pueden alcanzar, explorando aspectos de la realidad que nunca hemos conocido y que quizá nunca conozcamos directamente. Cuando un ordenador que se entrena solo elabora una estrategia de ajedrez que nunca se le ha ocurrido a ningún humano en la milenaria historia del juego, ¿qué ha descubierto y cómo lo ha descubierto? ¿Qué aspecto esencial

del juego, hasta ahora desconocido para las mentes humanas, ha percibido? Cuando un programa informático diseñado por humanos, que lleva a cabo un objetivo asignado por sus programadores —corregir errores en el software o perfeccionar los mecanismos de los vehículos autoconducidos— aprende y aplica un modelo que ningún humano reconoce o podría entender, ¿estamos avanzando hacia el conocimiento? ¿O el conocimiento se aleja de nosotros?

La humanidad ha experimentado cambios tecnológicos a lo largo de la historia. Sin embargo, solo en contadas ocasiones la tecnología ha transformado fundamentalmente la estructura social y política de nuestras sociedades. Lo más frecuente es que los marcos preexistentes a través de los cuales ordenamos nuestro mundo social se adapten y absorban la nueva tecnología, evolucionando e innovando dentro de categorías reconocibles. El automóvil sustituyó al caballo sin forzar un cambio total de la estructura social. El fusil sustituyó al mosquete, pero el paradigma general de la actividad militar convencional permaneció prácticamente inalterado. Solo en contadas ocasiones nos hemos encontrado con una tecnología que haya puesto en tela de juicio nuestros modos predominantes de explicar y ordenar el mundo. Pero la IA promete transformar todos los ámbitos de la experiencia humana. Y el núcleo de sus transformaciones se producirá en última instancia a nivel filosófico, transformando la forma en que los humanos entendemos la realidad y nuestro papel en ella.

La naturaleza sin precedentes de este proceso es tan profunda como desconcertante; habiendo entrado en él gradualmente, lo estamos experimentando pasivamente, en gran medida inconscientes de lo que ha hecho y de lo que probablemente hará en los próximos años. Sus cimientos los pusieron los ordenadores e Internet. Su apogeo será la IA omnipresente, que aumentará el pensamiento y la acción humanos de formas evidentes (como los nuevos fármacos y las traducciones automáticas de

idiomas) y menos conscientes (como los procesos de software que aprenden de nuestros movimientos y elecciones y se ajustan para anticipar o dar forma a nuestras necesidades futuras). Ahora que se ha demostrado lo prometedor de la IA y del aprendizaje automático, y que la potencia de cálculo necesaria para hacer funcionar una IA sofisticada está cada vez más disponible, pocos campos quedarán inalterados.

De forma persistente, a menudo imperceptible, pero ahora inevitable, una red de procesos de software se está desplegando por todo el mundo, impulsando y percibiendo el ritmo y el alcance de los acontecimientos, superponiéndose a aspectos de nuestra vida cotidiana —hogares, transporte, distribución de noticias, mercados financieros, operaciones militares— que antes nuestras mentes recorrían solas. A medida que más software incorpore la IA, y con el tiempo opere de formas que los humanos no crearon directamente o no puedan comprender del todo, será un aumentador dinámico de procesamiento de información de nuestras capacidades y experiencias, tanto dando forma a nuestras acciones como aprendiendo de ellas. A menudo, seremos conscientes de que esos programas nos están ayudando de una forma que nosotros no pretendíamos. Sin embargo, es posible que en un momento dado no sepamos exactamente qué están haciendo o identificando o por qué funcionan. La tecnología impulsada por IA se convertirá en un compañero permanente en la percepción y el procesamiento de la información, aunque ocupe un plano «mental» diferente al de los humanos. Tanto si la consideramos una herramienta, un socio o un rival, alterará nuestra experiencia como seres razonadores y cambiará de forma permanente nuestra relación con la realidad.

El viaje de la mente humana hasta el escenario central de la historia duró muchos siglos. En Occidente, la aparición de la imprenta y la Reforma protestante pusieron en tela de juicio las jerarquías oficiales y modificaron el marco de referencia de la sociedad: de la búsqueda del conocimiento de lo divino a través

de las Escrituras y su interpretación oficial se pasó a la búsqueda del conocimiento y la plenitud a través del análisis y la exploración individuales. El Renacimiento fue testigo del redescubrimiento de los escritos clásicos y los modos de investigación que se utilizaron para dar sentido a un mundo cuyos horizontes se expandían a través de la exploración global. Durante la Ilustración, la máxima de René Descartes, *Cogito ergo sum* (Pienso, luego existo), consagró la mente razonadora como capacidad definitoria de la humanidad y reivindicación de centralidad histórica. Esta noción también comunicaba el sentido de posibilidad engendrado por la ruptura del monopolio establecido sobre la información, que estaba en gran parte en manos de la Iglesia.

Ahora, el fin parcial de la postulada superioridad de la razón humana, junto con la proliferación de máquinas capaces de igualar o superar la inteligencia humana, promete transformaciones potencialmente más profundas incluso que las de la Ilustración. Aun cuando los avances en IA no producen inteligencia general artificial (AGI) —es decir, software capaz de realizar cualquier tarea intelectual a nivel humano y capaz de relacionar tareas y conceptos con otros a través de disciplinas—, el advenimiento de la IA alterará el concepto que la humanidad tiene de la realidad y, por tanto, de sí misma. Estamos avanzando hacia grandes logros, pero esos logros deberían suscitar una reflexión filosófica. Cuatro siglos después de que Descartes promulgara su máxima, se cierne una pregunta: si la IA «piensa», o se aproxima al pensamiento, ¿quiénes somos nosotros?

La IA marcará el comienzo de un mundo en el que las decisiones se tomarán principalmente de tres formas: por humanos (lo cual es familiar), por máquinas (lo cual se está volviendo familiar) y mediante la colaboración entre humanos y máquinas (lo cual no solo es desconocido, sino que no tiene precedentes). La IA también está transformando las máquinas, que hasta ahora eran nuestras herramientas, en nuestros socios. Empezaremos a

dar a la IA menos instrucciones específicas sobre cómo alcanzar exactamente los objetivos que le asignemos. Con mucha más frecuencia, presentaremos a la IA objetivos ambiguos y le preguntaremos: «¿Cómo, basándonos en tus conclusiones, deberíamos proceder?».

Este cambio no es ni intrínsecamente amenazador ni intrínsecamente redentor. Sin embargo, es tan diferente que muy probablemente alterará la trayectoria de las sociedades y el curso de la historia. La continua integración de la IA en nuestras vidas traerá consigo un mundo en el que se alcanzarán objetivos humanos aparentemente imposibles y en el cual los logros que antes se presumían exclusivamente humanos —escribir una canción, descubrir un tratamiento médico— serán generados por máquinas o en colaboración con ellas. Esta evolución transformará campos enteros al envolverlos en procesos asistidos por la IA, y a veces resultará difícil definir la línea divisoria entre la toma de decisiones puramente humana, la puramente basada en la IA y la híbrida entre humanos e IA.

En el ámbito político, el mundo está entrando en una era en la que los sistemas de IA basados en macrodatos están influyendo en aspectos cada vez más importantes: el diseño de mensajes políticos; la adaptación y distribución de esos mensajes a diversos grupos demográficos; la elaboración y aplicación de desinformación por parte de agentes malintencionados que pretenden sembrar la discordia social coma en vez de punto y coma y el diseño y despliegue de algoritmos para detectar, identificar y contrarrestar la desinformación y otras formas de datos perjudiciales. A medida que crece el papel de la IA en la definición y configuración del «espacio de la información», se hace más difícil anticipar dicho papel. En este espacio, como en otros, la IA opera a veces de formas que incluso sus diseñadores solo pueden elaborar en términos generales. Como resultado, las perspectivas de una sociedad libre, incluso del libre albedrío, pueden verse alteradas. Incluso si estas evoluciones resultan ser

benignas o reversibles, corresponde a las sociedades de todo el mundo comprender estos cambios para poder conciliarlos con sus valores, estructuras y contratos sociales.

Las instituciones y mandos de defensa se enfrentan a evoluciones no menos profundas. Cuando múltiples ejércitos adopten estrategias y tácticas configuradas por máquinas que perciben patrones que los soldados y estrategias humanos no pueden percibir, los equilibrios de poder se verán alterados y serán potencialmente más difíciles de calcular. Si se autoriza a esas máquinas a tomar decisiones autónomas sobre los objetivos, los conceptos tradicionales de defensa y disuasión —y las leyes de la guerra en su conjunto— pueden deteriorarse o, como mínimo, requerir una adaptación.

En tales casos, aparecerán nuevas divisiones dentro de las sociedades y entre ellas, entre quienes adopten la nueva tecnología y quienes opten por no utilizarla o carezcan de medios para desarrollar o adquirir algunas de sus aplicaciones. Cuando varios grupos o naciones adopten conceptos o aplicaciones diferentes de la IA, sus experiencias de la realidad pueden divergir de formas difíciles de predecir o salvar. A medida que las sociedades desarrollan sus propias asociaciones hombre-máquina —con objetivos distintos, modelos de entrenamiento diferentes y límites operativos y morales potencialmente incompatibles con respecto a la IA—, pueden derivar en rivalidad, incompatibilidad técnica y una incomprensión mutua cada vez mayor. La tecnología que inicialmente se creyó un instrumento para la trascendencia de las diferencias nacionales y la dispersión de la verdad objetiva puede, con el tiempo, convertirse en el método por el que las civilizaciones y los individuos divergen en realidades diferentes y mutuamente ininteligibles.

AlphaZero es un ejemplo ilustrativo. Demostró que la IA, al menos en los juegos, ya no estaba constreñida por los límites del conocimiento humano establecido. Es cierto que el tipo de IA que subyace en AlphaZero —aprendizaje automático en el que

los algoritmos se entrenan en redes neuronales profundas—tiene sus propias limitaciones. Pero en un número cada vez mayor de aplicaciones, las máquinas están ideando soluciones que parecen ir más allá del alcance de la imaginación humana. En 2016, una subdivisión de DeepMind, DeepMind Applied, desarrolló una IA (que funcionaba con muchos de los mismos principios que AlphaZero) para optimizar la refrigeración de los centros de datos de Google, sensibles a la temperatura. Aunque algunos de los mejores ingenieros del mundo ya habían abordado el problema, el programa de IA de DeepMind optimizó aún más la refrigeración y redujo el gasto de energía en un 40 % adicional, una mejora masiva sobre el rendimiento humano.⁶ Cuando la IA se aplique para lograr avances comparables en diversos campos de actividad, el mundo cambiará inevitablemente. Los resultados no serán simplemente formas más eficientes de realizar tareas humanas: en muchos casos, la IA sugerirá nuevas soluciones o direcciones que llevarán el sello de otra forma no humana de aprendizaje y evaluación lógica.

Una vez que el rendimiento de la IA supera al de los humanos en una tarea determinada, no aplicar esa IA, al menos como complemento de los esfuerzos humanos, puede parecer cada vez más perverso o incluso negligente. El hecho de que a un jugador de ajedrez asistido por IA se le aconseje sacrificar una pieza valiosa que los jugadores más sofisticados habían considerado tradicionalmente indispensable no tiene mayor importancia, pero en el contexto de la seguridad nacional, ¿qué ocurriría si la IA recomendará a un comandante en jefe sacrificar a un número significativo de ciudadanos o sus intereses, para salvar, según los cálculos y la valoración de la IA, a un número aún mayor? ¿Sobre qué base podría ignorarse ese sacrificio? ¿Estaría justificado? ¿Sabrán siempre los humanos qué cálculos ha hecho la IA? ¿Serán capaces los humanos de detectar elecciones (de la IA) no deseadas o de anularlas a tiempo? Si no somos capaces de comprender la lógica de cada decisión individual,

¿deberíamos aplicar sus recomendaciones solo por fe? Si no lo hacemos, ¿corremos el riesgo de interrumpir un rendimiento superior al nuestro? Incluso si somos capaces de comprender la lógica, el precio y el impacto de determinadas alternativas, ¿qué ocurre si nuestro adversario confía igualmente en la IA? ¿Cómo se logrará el equilibrio entre estas consideraciones o, en su caso, se reivindicará?

Tanto en el éxito de AlphaZero como en el descubrimiento de la halicina, la IA dependía de los humanos para definir el problema que resolvía. El objetivo de AlphaZero era ganar al ajedrez siguiendo las reglas del juego. El objetivo de la IA descubridora de la halicina era matar tantos patógenos como fuera posible: cuantos más patógenos matara sin dañar al huésped, más éxito tendría. Además, se designó como objetivo un reino más allá del alcance humano: en lugar de localizar vías conocidas de administración de fármacos, se le ordenó que buscara enfoques no descubiertos. La IA tuvo éxito porque el antibiótico que descubrió mataba a los patógenos. Pero fue especialmente innovador porque amplía las opciones de tratamiento al añadir un nuevo (y potente) antibiótico administrado a través de un nuevo mecanismo.

Está surgiendo una nueva asociación entre el hombre y la máquina: En primer lugar, los humanos definen un problema o un objetivo para una máquina. A continuación, una máquina, que opera en un ámbito fuera del alcance humano, determina el proceso óptimo a seguir. Una vez que una máquina ha llevado un proceso al ámbito humano, podemos intentar estudiarlo, comprenderlo e, idealmente, incorporarlo a la práctica existente. Desde la victoria de AlphaZero, su estrategia y táctica se han incorporado al juego humano, ampliando las concepciones humanas del ajedrez. Las Fuerzas Aéreas de EE. UU. han adaptado los principios subyacentes de AlphaZero a una nueva IA, μ Zero, que dirigió con éxito un avión de vigilancia U-2 en un vuelo de prueba: el primer programa informático que

pilota un avión militar y maneja sus sistemas de radar de forma autónoma, sin supervisión humana directa.⁷ La IA que descubrió la halicina ha ampliado los conceptos de los investigadores humanos, tanto en sentido estricto (erradicación de bacterias, administración de fármacos), como en sentido amplio (enfermedad, medicina, salud).

El hecho de que la actual colaboración hombre-máquina requiera un problema definible y un objetivo mensurable no es motivo para temer a las máquinas que todo lo saben y todo lo controlan. Sin embargo, la colaboración entre el hombre y la máquina supone un profundo cambio con respecto a la experiencia anterior.

Los motores de búsqueda plantean otro reto: hace diez años, cuando los motores de búsqueda funcionaban con minería de datos (y no con aprendizaje automático), si una persona buscaba «restaurantes gourmet» y luego «ropa», su búsqueda de lo segundo sería independiente de su búsqueda de lo primero. En ambos casos, un motor de búsqueda agregaba toda la información posible y ofrecía opciones al usuario, algo así como una guía telefónica digital o un catálogo temático. Pero los buscadores actuales se guían por modelos basados en el comportamiento humano observado. Si una persona busca «restaurantes gourmet» y luego «ropa», es posible que le aparezca ropa de diseño en lugar de alternativas más asequibles. La ropa de diseño puede ser lo que el buscador busca. Pero hay una diferencia entre elegir entre un abanico de opciones y emprender una acción —en este caso, hacer una compra; en otros, adoptar una postura o ideología política o filosófica— sin saber nunca cuál era el abanico inicial de posibilidades o implicaciones, confiando a una máquina la configuración preventiva de las opciones.

Hasta ahora, la elección basada en la razón ha sido la prerrogativa —y, desde la Ilustración, el atributo definitorio— de la humanidad. La llegada de máquinas capaces de aproximarse a la

razón humana alterará tanto a los humanos como a las máquinas. Las máquinas iluminarán a los humanos, ampliando nuestra realidad de formas que no esperábamos ni pretendíamos necesariamente provocar (también será posible lo contrario: que las máquinas que consumen el conocimiento humano se utilicen para disminuirnos). Simultáneamente, los humanos crearán máquinas capaces de realizar descubrimientos y conclusiones sorprendentes, capaces de aprender y evaluar el significado de sus descubrimientos. El resultado será una nueva época.

La humanidad tiene siglos de experiencia en el uso de máquinas para aumentar, automatizar y, en muchos casos, sustituir el trabajo manual. Las olas de cambio que trajo consigo la Revolución Industrial aún resuenan en los ámbitos de la economía, la política, la vida intelectual y los asuntos internacionales. Al no reconocer las muchas comodidades modernas que ya proporciona la IA, hemos llegado a depender de la tecnología de forma lenta, casi pasiva, sin registrar ni el hecho de nuestra dependencia ni las implicaciones de esta. En la vida cotidiana, la IA es nuestra compañera, ayudándonos a tomar decisiones sobre qué comer, qué vestir, en qué creer, adónde ir y cómo llegar.

Aunque la IA puede sacar conclusiones, hacer predicciones y tomar decisiones, no posee autoconciencia, es decir, la capacidad de reflexionar sobre su papel en el mundo. No tiene intención, motivación, moralidad ni emoción; incluso sin estos atributos, es probable que desarrolle medios diferentes y no intencionados de alcanzar los objetivos asignados. Pero, inevitablemente, cambiará a los seres humanos y los entornos en los que viven. Cuando los individuos crecen o se entrenan con él, pueden sentir la tentación, incluso inconsciente, de antropomorfizarlo y tratarlo como a un semejante.

Aunque la tecnología parece opaca y misteriosa para la inmensa mayoría de la población humana, cada vez más personas de universidades, empresas y gobiernos han aprendido

a construir, utilizar y desplegar la IA en productos de consumo habituales, a través de los cuales muchos de nosotros ya estamos interactuando con ellos, a sabiendas o no. Pero mientras crece el número de individuos capaces de crear IA, las filas de quienes contemplan las implicaciones de esta tecnología para la humanidad —sociales, legales, filosóficas, espirituales, morales— siguen siendo peligrosamente escasas.

Con la ayuda del avance y el uso cada vez mayor de la IA, la mente humana está accediendo a nuevas perspectivas, poniendo a la vista objetivos antes inalcanzables. Por ejemplo, modelos con los que predecir y mitigar catástrofes naturales, un conocimiento más profundo de las matemáticas y una comprensión más completa del universo y de la realidad en la que reside. Pero estas y otras posibilidades se están adquiriendo —en gran medida sin fanfarria— alterando la relación humana con la razón y la realidad. Se trata de una revolución para la que los conceptos filosóficos y las instituciones sociales existentes nos dejan muy poco preparados.