

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*



Henry A. Kissinger
Eric Schmidt
Daniel Huttenlocher

LA ERA DE LA INTELIGENCIA ARTIFICIAL

y nuestro futuro humano

Henry A. Kissinger
Eric Schmidt
Daniel Huttenlocher



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

Agradecimientos¹

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.

¹ *N. de la T.*: En aquellos casos en que el libro citado tiene una edición española, se ha sustituido la referencia en inglés por la de la edición disponible en España, para facilitar su búsqueda y consulta. Igualmente, se han traducido aquellas notas que constituyen comentarios de los autores. Obviamente, las referencias de internet y de artículos, disponibles solamente en inglés, se han mantenido en su idioma original para permitir su localización.

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](https://www.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

Prefacio

1. Dónde estamos
2. Cómo hemos llegado hasta aquí
3. De Turing a nuestros días
4. Plataformas de redes globales
5. Seguridad y orden mundial
6. IA e identidad humana
7. La IA y el futuro

Epílogo: Una nueva realidad

Créditos

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.

¹ «AI Startups Raised USD734bn in Total Funding in 2020», *Equity Wire*, November 19, 2020, <https://www.privateequitywire.co.uk/2020/11/19/292458/ai-startups-raised-usd734bn-total-funding-2020>.

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; como 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 recomendara 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.

¹ Mike Klein, «Google's AlphaZero Destroys Stockfish in 100-Game Match», *Chess.com*, December 6, 2017, <https://www.chess.com/news/view/google-s-alphazero-destroys-stockfish-in-100-game-match>; <https://perma.cc/8WGK-HKYZ>; Pete, «AlphaZero Crushes Stockfish in New 1,000-Game Match», *Chess.com*, April 17, 2019, <https://www.chess.com/news/view/updated-alphazero-crushes-stockfish-in-new-1-000-game-match>.

² Garry Kasparov. Prólogo. *Game Changer: AlphaZero's Groundbreaking Chess Strategies and the Promise of AI* por Matthew Sadler y Natasha Regan, *New in Chess*, 2019, 10.

³ «Step 1: Discovery and Development,» US Food and Drug Administration, January 4, 2018, <https://www.fda.gov/patients/drug-development-process/step-1-discovery-and-development>.

⁴ Jo Marchant, «Powerful Antibiotics Discovered Using AI», *Nature*, February 20, 2020, <https://www.nature.com/articles/d41586-020-00018-3>.

⁵ Raphaël Millière (@raphamilliere), «I asked GPT-3 to write a response to the philosophical essays written about it...». July 30, 2020, 5:24 a.m., <https://twitter.com/raphamilliere/status/1289129723310886912/photo/1>; Justin Weinberg, «Update: Some Replies by GPT-3,» *Daily Nous*, July 30, 2020, <https://dailynous.com/2020/07/30/philosophers-gpt-3/#gpt3replies>.

⁶ Richard Evans y Jim Gao, «DeepMind AI Reduces Google Data Centre Cooling Bill by 40%», *DeepMind blog*, July 20, 2016, <https://deepmind.com/blog/article/deepmind-ai-reduces-google-data-centre-cooling-bill-40>.

⁷ Will Roper, «AI Just Controlled a Military Plane for the First Time Ever», Popular Mechanics, December 16, 2020, <https://www.popularmechanics.com/military/aviation/a34978872/artificial-intelligence-controls-u2-spy-plane-air-force-exclusive>.

Cómo hemos llegado hasta aquí

Tecnología y pensamiento humano

A lo largo de la historia, los seres humanos hemos luchado por comprender plenamente aspectos de nuestra experiencia y de los entornos en los que vivimos. Cada sociedad, a su manera, ha indagado en la naturaleza de la realidad: ¿cómo entenderla? ¿Predecirla? ¿Configurarla? ¿Moderarla? A medida que se han ido planteando estas cuestiones, cada sociedad ha llegado a su propio conjunto de acuerdos con el mundo. En el centro de estas adaptaciones ha estado el concepto de la relación de la mente humana con la realidad: su capacidad de conocer su entorno, de realizarse mediante el conocimiento y, al mismo tiempo, de estar inherentemente limitada por él. Incluso si una época o una cultura consideraba que la razón humana era limitada —incapaz de percibir o comprender la vasta extensión del universo o las dimensiones esotéricas de la realidad—, al ser humano razonador individual se le ha concedido un orgulloso lugar como el ser terrenal más capaz de comprender y dar forma al mundo. El ser humano ha respondido a su entorno y se ha reconciliado con él identificando fenómenos que puede estudiar y, en última instancia, explicar, ya sea científicamente, teológicamente o de ambas formas. Con la llegada de la IA, la humanidad está creando un nuevo y poderoso actor en esta búsqueda. Para comprender la importancia de esta evolución, haremos un breve repaso del camino por el cual la razón humana, a

través de sucesivas épocas históricas, ha adquirido su estimado estatus.

Cada época histórica se ha caracterizado por un conjunto de explicaciones interrelacionadas de la realidad y de las disposiciones sociales, políticas y económicas basadas en ellas. El mundo clásico, la Edad Media, el Renacimiento y el mundo moderno cultivaron sus conceptos del individuo y la sociedad, teorizando sobre dónde y cómo encaja cada uno en el orden perdurable de las cosas. Cuando las concepciones predominantes ya no bastaban para explicar las percepciones de la realidad —acontecimientos vividos, descubrimientos realizados, otras culturas encontradas— se producían revoluciones en el pensamiento (y a veces en la política) y nacía una nueva época. La incipiente era de la inteligencia artificial plantea cada vez más desafíos epocales al concepto actual de la realidad.

En Occidente, el respeto central de la razón se originó en la Grecia y Roma antiguas. Estas sociedades elevaron la búsqueda del conocimiento a un aspecto definitorio, tanto de la realización individual como del bien colectivo. En la República de Platón, la famosa alegoría de la caverna habla de la centralidad de la búsqueda. Construida como un diálogo entre Sócrates y Glaucón, la alegoría compara a la humanidad con un grupo de prisioneros encadenados a la pared de una caverna. Al ver las sombras proyectadas en la pared de la cueva desde la boca iluminada por el sol, los prisioneros creen que tales sombras son la realidad. Según Sócrates, el filósofo es semejante al prisionero que se libera, asciende a la superficie y percibe la realidad a plena luz del día. Del mismo modo, la búsqueda platónica para vislumbrar la verdadera forma de las cosas suponía la existencia de una realidad objetiva —de hecho, ideal— hacia la cual la humanidad tiene la capacidad de caminar, aunque nunca llegue a alcanzarla del todo.

La convicción de que lo que vemos refleja la realidad —y de que podemos comprender plenamente al menos aspectos de esta realidad utilizando la disciplina y la razón— inspiró a los filósofos

griegos y a sus herederos a alcanzar grandes logros. Pitágoras y sus discípulos exploraron la conexión entre las matemáticas y las armonías internas de la naturaleza, y elevaron esta búsqueda a doctrina espiritual esotérica. Tales de Mileto estableció un método de investigación comparable al método científico moderno, que acabó inspirando a los pioneros de la ciencia moderna. La amplia clasificación del conocimiento de Aristóteles, la geografía pionera de Ptolomeo y *Sobre la naturaleza de las cosas*, de Lucrecio, reflejan una confianza esencial en la capacidad de la mente humana para descubrir y comprender, al menos, aspectos sustanciales del mundo. Estas obras y el estilo de lógica que empleaban se convirtieron en vehículos educativos que permitieron a los eruditos desarrollar inventos, aumentar las defensas y diseñar y construir grandes ciudades que, a su vez, se convirtieron en centros de aprendizaje, comercio y exploración exterior.

Aun así, el mundo clásico percibía fenómenos aparentemente inexplicables para los que no se podían encontrar explicaciones adecuadas únicamente en la razón. Estas experiencias misteriosas se atribuían a una serie de dioses que solo los devotos e iniciados podían conocer simbólicamente, y cuyos ritos y rituales concomitantes únicamente los devotos e iniciados podían observar. Al relatar los logros del mundo clásico y el declive del Imperio Romano a través de su propia lente ilustrada, el historiador del siglo XVIII Edward Gibbon describió un mundo en el que las deidades paganas se erigían en explicaciones de fenómenos naturales fundamentalmente misteriosos que se consideraban importantes o amenazadores:

La fina textura de la mitología pagana estaba entretejida con materiales diversos pero no discordantes... Las deidades de mil arboledas y mil arroyos poseían, en paz, su influencia local y respectiva; no podía el romano que menospreciaba la ira del Tíber, burlarse del egipcio que presentaba su ofrenda al genio benéfico del Nilo. Los poderes visibles de la Naturaleza, los planetas y los elementos, eran los mismos en todo el universo.

Los gobernantes invisibles del mundo moral se fundieron inevitablemente en un molde similar de ficción y alegoría.¹

Aún no se sabía científicamente por qué cambiaban las estaciones, por qué la Tierra parecía morir y volver a la vida a intervalos regulares. Las culturas griega y romana reconocían los patrones temporales de los días y los meses, pero no habían llegado a una explicación deducible únicamente mediante la experimentación o la lógica. Por eso se ofrecieron como alternativa los famosos Misterios de Eleusis, que representaban el drama de la diosa de la cosecha, Deméter, y su hija, Perséfone, condenadas a pasar una parte del año en el frío inframundo del Hades. A través de estos ritos esotéricos, los participantes «conocían» la realidad más profunda de las estaciones: la abundancia o escasez agrícola de la región y su impacto en la sociedad. Del mismo modo, un comerciante que emprendiera un viaje podría adquirir un concepto básico de las mareas y la geografía marítima a través del conocimiento práctico acumulado de su comunidad; no obstante, seguiría intentando contentar a las deidades del mar, y a las de los viajes seguros de ida y vuelta, quienes creía controlaban los medios y fenómenos por los que pasaría.

El auge de las religiones monoteístas desequilibró la mezcla de razón y fe que durante tanto tiempo había dominado la búsqueda clásica del conocimiento del mundo. Aunque los filósofos clásicos habían reflexionado tanto sobre la naturaleza de la divinidad como sobre la divinidad de la naturaleza, rara vez habían postulado una única figura o motivación subyacente capaz de ser nombrada o venerada definitivamente. Sin embargo, para la Iglesia primitiva, estas exploraciones discursivas de las causas y los misterios eran otros tantos callejones sin salida o, según las valoraciones más caritativas o pragmáticas, precursores insólitos de la revelación de la sabiduría cristiana. La realidad oculta que el mundo clásico se había esforzado por percibir se consideraba divina, accesible solo parcial e indirectamente a través del culto. Este proceso fue mediado por una

institución religiosa que durante siglos mantuvo prácticamente el monopolio de la investigación académica, guiando a los individuos a través de los sacramentos hacia una comprensión de las Escrituras, redactadas y predicadas en un lenguaje que pocos profanos entendían.

La recompensa prometida a los individuos que seguían la fe «correcta» y se adherían a este camino hacia la sabiduría era la admisión en una vida después de la muerte, un plano de existencia considerado más real y significativo que la realidad observable. En la Edad Media —el periodo que va desde la caída de Roma, en el siglo v, hasta la conquista de Constantinopla por el Imperio Turco Otomano, en el siglo xv— la humanidad, al menos en Occidente, buscaba conocer primero a Dios y después al mundo. El mundo solo podía conocerse a través de Dios; la teología filtraba y ordenaba las experiencias individuales de los fenómenos naturales que se presentaban ante ellos. Cuando los primeros pensadores y científicos modernos, como Galileo, empezaron a explorar el mundo directamente, modificando sus explicaciones a la luz de la observación científica, fueron castigados y perseguidos por atreverse a omitir la teología como intermediaria.

Durante la época medieval, la escolástica se convirtió en la guía principal para la búsqueda permanente de la comprensión de la realidad percibida, mediante la veneración de la relación entre la fe, la razón y la Iglesia, que seguía siendo el árbitro de la legitimidad cuando se trataba de creencias y (al menos en teoría) de la legitimidad de los líderes políticos. Aunque la creencia generalizada era que la Cristiandad debía estar unificada, tanto teológica como políticamente, la realidad desmentía esta aspiración; desde el principio, hubo un enfrentamiento entre una variedad de sectas y unidades políticas. Sin embargo, a pesar de esta práctica, la cosmovisión europea no se actualizó durante muchas décadas. Se hicieron enormes progresos en la descripción y representación del universo: el periodo produjo la teología de Santo Tomás de Aquino, la poesía de Geoffrey Chaucer, la pintura de Giotto di Bondone y la

exploración de Marco Polo. En cambio, se avanzó mucho menos en la explicación del mundo. Todo fenómeno desconcertante, grande o pequeño, se atribuía a la obra del Señor.

En los siglos xv y xvi, el mundo occidental experimentó dos revoluciones gemelas que introdujeron una nueva época y, con ella, un nuevo concepto del papel de la mente y la conciencia humanas individuales en la navegación por la realidad. La invención de la imprenta posibilitó la circulación directa de materiales e ideas entre grandes grupos de personas en lenguas que entendían y no en el latín de las clases eruditas, y anuló la confianza histórica de la gente en que la Iglesia interpretara conceptos y creencias por ellos. Con la ayuda de la tecnología, los líderes de la Reforma protestante declararon que los individuos eran capaces —de hecho, responsables— de definir lo divino por sí mismos.

Al dividir el mundo cristiano, la Reforma validó la posibilidad de que la fe individual existiera independiente del arbitraje eclesiástico. A partir de ese momento, la autoridad —en la religión y, con el tiempo, en otros ámbitos— se sometió a la prueba de la investigación autónoma.

Durante esta época revolucionaria, la tecnología innovadora, los paradigmas novedosos y las adaptaciones políticas y sociales generalizadas se reforzaron mutuamente. Una vez que un libro pudo imprimirse y distribuirse fácilmente con una sola máquina y un solo operario —sin el trabajo costoso y especializado de los copistas monásticos—, las nuevas ideas pudieron difundirse y amplificarse más rápido de lo que podían restringirse. Las autoridades centralizadas —ya fuera la Iglesia Católica, el Sacro Imperio Romano Germánico dirigido por los Habsburgo (el sucesor teórico del dominio unificado de Roma sobre el continente europeo) o los gobiernos nacionales y locales— ya no podían detener la proliferación de la tecnología de impresión ni prohibir eficazmente las ideas desfavorables. Como Londres, Ámsterdam y otras ciudades importantes se negaron a proscribir la difusión de material impreso, los librepensadores que habían sido acosados por sus gobiernos

nacionales pudieron encontrar refugio y acceso a industrias editoriales avanzadas en sociedades cercanas. La visión de unidad doctrinal, filosófica y política dio paso a la diversidad y la fragmentación, en muchos casos acompañada del derrocamiento de las clases sociales establecidas y de violentos conflictos entre las facciones contendientes. Una época definida por un extraordinario progreso científico e intelectual se vio emparejada con disputas religiosas, dinásticas, nacionales y de clase casi constantes que provocaron continuos trastornos y pusieron en peligro la vida y el sustento de las personas.

A medida que la autoridad intelectual y política se fragmentaba en medio de la efervescencia doctrinal, se producían exploraciones artísticas y científicas de notable riqueza, en parte mediante la recuperación de textos, modos de aprendizaje y argumentación clásicos. Durante este *Renaissance*, o renacimiento, del aprendizaje clásico, las sociedades produjeron arte, arquitectura y filosofía que buscaban simultáneamente celebrar los logros humanos e inspirarlos aún más. El humanismo, principio rector de la época, pretendía fomentar individuos capaces de participar plenamente en la vida cívica a través de un pensamiento y una expresión claros. Según el humanismo, estas virtudes se cultivaban a través de las humanidades: arte, escritura, retórica, historia, política y filosofía. En consecuencia, los hombres del Renacimiento que dominaban estos campos — Leonardo da Vinci, Miguel Ángel, Rafael — llegaron a ser venerados. Ampliamente adoptado, el humanismo cultivó el amor por la lectura y el aprendizaje, y la primera facilitó el segundo.

El redescubrimiento de la ciencia y la filosofía griegas inspiró nuevas investigaciones sobre los mecanismos subyacentes del mundo natural y los medios para medirlos y catalogarlos. Cambios análogos comenzaron a producirse en el ámbito de la política y el Estado. Los eruditos se atrevieron a formar sistemas de pensamiento basados en principios organizativos que iban más allá de la restauración de la unidad cristiana continental bajo la égida moral del Papa. El diplomático y filósofo italiano Nicolás Maquiavelo,

también clasicista, defendió que los intereses del Estado eran distintos de su relación con la moral cristiana y se esforzó por esbozar principios racionales, aunque no siempre atractivos, que permitieran perseguir dichos intereses.²

Esta exploración del conocimiento histórico y la creciente sensación de poder sobre los mecanismos de la sociedad también inspiraron una era de exploración geográfica, en la que el mundo occidental se expandió y encontró nuevas sociedades, formas de creencia y tipos de organización política. Las sociedades más avanzadas y las mentes más eruditas de Europa se enfrentaron de repente a un nuevo aspecto de la realidad: sociedades con dioses diferentes, historias divergentes y, en muchos casos, formas propias e independientes de logros económicos y complejidad social. Para la mente occidental, formada en la convicción de su propia centralidad, estas sociedades organizadas de forma independiente planteaban profundos retos filosóficos. Culturas separadas con fundamentos distintos y sin conocimiento de las escrituras cristianas habían desarrollado existencias paralelas, sin conocimiento aparente de (o interés en) la civilización europea, que Occidente había asumido como el pináculo evidente de los logros humanos. En algunos casos, como en los encuentros de los conquistadores españoles con el Imperio azteca en México, las ceremonias religiosas indígenas y las estructuras políticas y sociales parecían comparables a las europeas.

Para los exploradores que se detuvieron en sus conquistas el tiempo suficiente para reflexionar sobre ellas, esta extraña correspondencia generó preguntas inquietantes: ¿Eran válidas de forma independiente las culturas y experiencias divergentes de la realidad? ¿Funcionaban las mentes y las almas de los europeos con arreglo a los mismos principios que las de los habitantes de América, China y otras tierras lejanas? ¿Estaban estas civilizaciones recién descubiertas esperando a que los europeos avalaran nuevos aspectos de la realidad —revelación divina, progreso científico— para despertar a la verdadera naturaleza de las cosas? ¿O siempre habían estado participando en la misma experiencia humana, respondiendo

a su propio entorno e historia y desarrollando sus propias adaptaciones paralelas a la realidad, cada una con sus puntos fuertes y logros relativos?

Aunque la mayoría de los exploradores y pensadores occidentales de la época llegaron a la conclusión de que estas sociedades recién descubiertas carecían de conocimientos fundamentales dignos de ser adoptados, las experiencias comenzaron a ampliar la apertura de la mente occidental. El horizonte se amplió para las civilizaciones de todo el planeta, lo cual les obligó a reconocer la amplitud y profundidad física y experiencial del mundo. En algunas sociedades occidentales, este proceso dio lugar a los conceptos de humanidad universal y derechos humanos, nociones que algunas de estas mismas sociedades acabaron promoviendo en periodos de reflexión posteriores.

Los avances tecnológicos y metodológicos, que incluían mejores lentes ópticas e instrumentos de medición más precisos, la manipulación química y el desarrollo de normas de investigación y observación luego denominadas «método científico», permitieron a los investigadores observar con mayor precisión los planetas y las estrellas, el comportamiento y la composición de las sustancias materiales y las minucias de la vida microscópica. Los científicos podían progresar de forma iterativa basándose tanto en sus observaciones personales como en las de sus colegas: cuando una teoría o predicción podía validarse empíricamente, se revelaban nuevos hechos que podían servir de punto de partida para nuevas preguntas. De este modo, salían a la luz nuevos descubrimientos, patrones y conexiones, muchos de los cuales podían aplicarse a aspectos prácticos de la vida cotidiana: determinar la hora, navegar por el océano, sintetizar compuestos útiles.

Tan acelerado fue el progreso en los siglos XVI y XVII —con asombrosos descubrimientos en matemáticas, astronomía y ciencias naturales— que provocó una especie de desorientación filosófica. Como la doctrina eclesiástica aún definía oficialmente los límites de las exploraciones intelectuales permisibles durante este periodo,

estos avances produjeron descubrimientos de una audacia considerable. La visión de Copérnico de un sistema heliocéntrico, las leyes del movimiento de Newton, la catalogación por van Leeuwenhoek de un mundo microscópico vivo... estos y otros avances provocaron la sensación general de que se estaban desvelando nuevas capas de la realidad. El resultado fue la incongruencia: las sociedades permanecían unidas en su monoteísmo, pero divididas por interpretaciones y exploraciones opuestas de la realidad. Necesitaban un concepto —de hecho, una filosofía— que guiara su búsqueda para comprender el mundo y su papel en él.

Los filósofos de la Ilustración respondieron a la llamada, declarando que la razón —el poder de comprender, pensar y juzgar— era tanto el método como el propósito de interactuar con el entorno. «Nuestra alma está hecha para pensar, es decir, para percibir», escribió el filósofo y polímata francés Montesquieu, «pero tal ser debe tener curiosidad, porque así como todas las cosas forman una cadena en la que cada idea precede a otra y sigue a otra, no se puede querer ver una sin desear ver la otra».³ La relación entre la primera pregunta de la humanidad (la naturaleza de la realidad) y la segunda (su papel en la realidad) se reforzó por sí misma: si la razón engendró la conciencia, cuanto más razonaban los humanos, más cumplían su propósito. Percibir y elaborar el mundo era el proyecto más importante en el que estaban o estarían comprometidos. Había nacido la edad de la razón.

En cierto sentido, Occidente volvía a muchas de las cuestiones fundamentales con las que ya habían luchado los antiguos griegos: ¿Qué es la realidad? ¿Qué es lo que la gente quiere conocer y experimentar, y cómo lo sabrá cuando se encuentre con ello? ¿Puede el ser humano percibir la realidad en sí misma en oposición a sus pensamientos? En caso afirmativo, ¿cómo? ¿Qué significa *ser* y *saber*? Libres de las ataduras de la tradición, o al menos creyendo justificado interpretarla de nuevo, estudiosos y filósofos volvieron a investigar estas cuestiones. Las mentes que emprendieron este viaje

estaban dispuestas a recorrer un camino precario, arriesgando las aparentes certezas de sus tradiciones culturales y sus concepciones establecidas de la realidad.

En esta atmósfera de desafíos intelectuales, conceptos antaño axiomáticos —la existencia de la realidad física, la naturaleza eterna de las verdades morales— se vieron repentinamente cuestionados.⁴ El *Tratado sobre los principios del conocimiento humano*, de 1710, del obispo Berkeley, sostenía que la realidad no consistía en objetos materiales, sino en Dios y en mentes cuya percepción de la realidad aparentemente sustantiva, según él, *eran* verdaderamente la realidad. Gottfried Wilhelm Leibniz, filósofo alemán de finales del siglo XVII y principios del XVIII, inventor de las primeras máquinas de calcular y pionero de algunos aspectos de la teoría informática moderna, defendió indirectamente un concepto tradicional de la fe al afirmar que las mónadas (unidades no reducibles a partes más pequeñas, con un papel intrínseco y divinamente designado en el universo) constituían la esencia subyacente de las cosas. El filósofo holandés del siglo XVII Baruch Spinoza, navegando por el plano de la razón abstracta con audacia y brillantez, trató de aplicar la lógica geométrica euclidiana a los preceptos éticos para «demostrar» un sistema ético en el que un Dios universal permitía y recompensaba la bondad humana. Esta filosofía moral no se basaba en escrituras ni milagros; Spinoza pretendía llegar al mismo sistema de verdades subyacente mediante la sola aplicación de la razón. En la cima del conocimiento humano, sostenía Spinoza, se encontraba la capacidad de la mente para razonar su camino hacia la contemplación de lo eterno: conocer «la idea de la mente misma» y reconocer, a través de la mente, al infinito y siempre presente «Dios como causa». Este conocimiento, sostenía Spinoza, era eterno, la forma última y de hecho perfecta de conocimiento. Lo llamó «el amor intelectual de Dios».⁵

Como resultado de estas exploraciones filosóficas pioneras, la relación entre la razón, la fe y la realidad se hizo cada vez más incierta. En esta brecha entró Immanuel Kant, un filósofo y profesor

alemán que trabajaba en la ciudad prusiana oriental de Königsberg.⁶ En 1781, Kant publicó su *Crítica de la razón pura*, una obra que ha inspirado y dejado perplejos a los lectores desde entonces. Alumno de los tradicionalistas y corresponsal de los racionalistas puros, Kant lamentó no estar de acuerdo con ninguno de ellos y trató de tender un puente entre las afirmaciones tradicionales y la nueva confianza de su época en el poder de la mente humana. En su *Crítica...*, Kant propuso que «la razón debería asumir de nuevo la más difícil de todas sus tareas, a saber, la del autoconocimiento».⁷ La razón, argumentaba Kant, debería aplicarse para comprender sus propias limitaciones.

Según el relato de Kant, la razón humana tiene la capacidad de conocer la realidad en profundidad, aunque a través de una lente inevitablemente imperfecta. La cognición y la experiencia humanas filtran, estructuran y distorsionan todo lo que conocemos, incluso cuando intentamos razonar «puramente» mediante la lógica. La realidad objetiva en sentido estricto —lo que Kant llamaba la cosa— en—sí— está siempre presente, pero intrínsecamente más allá de nuestro conocimiento directo. Kant postuló un reino de noúmeno, o «cosas tal y como son entendidas por el pensamiento puro», existentes independientemente de la experiencia o de la filtración a través de conceptos humanos. Sin embargo, Kant sostenía que, como la mente humana se basa en el pensamiento conceptual y en la experiencia vivida, nunca podría alcanzar el grado de pensamiento puro necesario para conocer esta esencia interior de las cosas.⁸ Podemos mantener creencias sobre lo que hay más allá y más adentro, pero esto no constituye un verdadero conocimiento de ello.⁹

Durante los doscientos años siguientes, la distinción esencial de Kant entre la cosa en sí y el mundo inevitablemente filtrado que experimentamos apenas pareció importar. Aunque la mente humana pudiera presentar una imagen imperfecta de la realidad, era la única imagen disponible. Lo que las estructuras de la mente humana impedían ver, presumiblemente quedaría prohibido para siempre, o

inspiraría la fe y la conciencia del infinito. Sin ningún mecanismo alternativo para acceder a la realidad, parecía que los puntos ciegos de la humanidad permanecerían ocultos. Si la percepción y la razón humanas debían ser la medida definitiva de las cosas, al carecer de alternativa, durante un tiempo llegaron a serlo. Pero la IA está empezando a proporcionar un medio alternativo de acceder —y, por tanto, de comprender— la realidad.

Durante las generaciones posteriores a Kant, la búsqueda del conocimiento de la cosa en sí adoptó dos formas: la observación cada vez más precisa de la realidad y la catalogación cada vez más amplia del conocimiento. Nuevos y vastos campos de fenómenos parecían cognoscibles, capaces de ser descubiertos y catalogados mediante la aplicación de la razón. A su vez, se creía que esos catálogos exhaustivos podrían desvelar lecciones y principios aplicables a las cuestiones científicas, económicas, sociales y políticas más apremiantes de la época. El esfuerzo más arrollador en este sentido fue la *Encyclopédie*, editada por el filósofo francés Denis Diderot. En veintiocho volúmenes (diecisiete de artículos y once de ilustraciones), 75 000 entradas y 18 000 páginas, la *Encyclopédie* de Diderot recogió los diversos hallazgos y observaciones de grandes pensadores en numerosas disciplinas, sus descubrimientos y deducciones y la relación entre los hechos y principios resultantes. Reconociendo el hecho de que su intento de catalogar todos los fenómenos de la realidad en un libro unificado era en sí mismo un fenómeno único, la enciclopedia incluyó una entrada autorreferencial sobre la palabra enciclopedia.

En el ámbito político, por supuesto, diversas mentes razonadoras (al servicio de diversos intereses estatales) no eran tan propensas a llegar a las mismas conclusiones. Federico el Grande de Prusia, un estadista prototípico de principios de la Ilustración, mantuvo correspondencia con Voltaire, entrenó a sus tropas a la perfección y se apoderó de la provincia de Silesia sin más advertencia ni justificación que el interés nacional de Prusia. Su ascenso provocó maniobras que desembocaron en la Guerra de los Siete Años, en

cierto sentido, la primera guerra mundial, porque se libró en tres continentes. Asimismo, la Revolución Francesa, uno de los movimientos políticos más orgullosamente «racionales» de la época, produjo trastornos sociales y violencia política a una escala nunca vista en Europa durante siglos. Al separar la razón de la tradición, la Ilustración produjo un nuevo fenómeno: la razón armada, fundida con las pasiones populares, reordenaba y arrasaba las estructuras sociales en nombre de conclusiones «científicas» sobre el rumbo de la historia. Las innovaciones posibilitadas por el método científico moderno aumentaron el poder destructivo de las armas y acabaron dando paso a la era de la guerra total: conflictos caracterizados por la movilización de la sociedad y la destrucción industrial.¹⁰

La Ilustración aplicó la razón tanto para tratar de definir sus problemas como para intentar resolverlos. A tal fin, el ensayo de Kant «La paz perpetua» postulaba (con cierto escepticismo) que la paz podría alcanzarse mediante la aplicación de normas consensuadas que rigieran las relaciones entre Estados independientes. Como aún no se habían establecido tales normas mutuas, al menos en una forma que los monarcas pudieran discernir o fuera probable que siguieran, Kant propuso un «artículo secreto de la paz perpetua», y sugirió que «los Estados que están armados para la guerra» consultaran «las máximas de los filósofos».¹¹ La visión de un sistema internacional razonado, negociado y sujeto a normas ha estado en el candelero desde entonces, con la contribución de filósofos y politólogos, pero con un éxito intermitente.

Movidos por las convulsiones políticas y sociales de la modernidad, los pensadores se mostraron cada vez más dispuestos a cuestionar si la percepción humana, ordenada por la razón humana, era la única métrica para dar sentido a la realidad. A finales del siglo XVIII y principios del XIX, el Romanticismo —que era una reacción a la Ilustración— consideraba el sentimiento y la imaginación humanos como verdaderos homólogos de la razón; exaltaba las tradiciones populares, la experiencia de la naturaleza y

una época medieval reimaginada como preferibles a las certezas mecanicistas de la era moderna.

Mientras tanto, la razón —en forma de física teórica avanzada— empezó a avanzar más hacia la cosa—en—sí de Kant, con consecuencias científicas y filosóficas desorientadoras. A finales del siglo XIX y principios del XX, los avances en las fronteras de la física empezaron a revelar aspectos inesperados de la realidad. El modelo clásico de la física, cuyos fundamentos se remontaban a la Ilustración, planteaba un mundo explicable en términos de espacio, tiempo, materia y energía, con propiedades absolutas y coherentes. Sin embargo, a medida que los científicos buscaban una explicación más clara de las propiedades de la luz, se encontraron con resultados que esta concepción tradicional no podía explicar. El brillante e iconoclasta físico teórico Albert Einstein resolvió muchos de estos enigmas con su trabajo pionero en física cuántica y sus teorías de la relatividad especial y general. Al hacerlo, reveló una imagen de la realidad física que parecía nuevamente misteriosa. El espacio y el tiempo se unían como un único fenómeno en el que las percepciones individuales no parecían estar sujetas a las leyes de la física clásica.¹²

Al desarrollar una mecánica cuántica para describir este sustrato de la realidad física, Werner Heisenberg y Niels Bohr pusieron en tela de juicio antiguas suposiciones sobre la naturaleza del conocimiento. Heisenberg subrayó la imposibilidad de evaluar simultáneamente y con precisión la posición y el momento de una partícula. Este «principio de incertidumbre» (como llegó a conocerse) implicaba que podría no disponerse de una imagen completamente exacta de la realidad en un momento dado. Además, Heisenberg sostenía que la realidad física no tenía una forma inherente independiente, sino que era creada por el proceso de observación: «Creo que se puede formular sucintamente la aparición de la “trayectoria” clásica de una partícula... la “trayectoria” solo se produce porque la observamos».¹³

La cuestión de si la realidad tiene una única forma verdadera y objetiva —y si las mentes humanas pueden acceder a ella— ha preocupado a los filósofos desde Platón. En obras como *Física y Filosofía: La revolución de la ciencia moderna* (1958), Heisenberg exploró la interacción entre ambas disciplinas y los misterios que la ciencia empezaba a desvelar. Bohr, en su propio trabajo pionero, subrayó que la observación afectaba y ordenaba la realidad. Según Bohr, el propio instrumento científico —durante mucho tiempo considerado una herramienta objetiva y neutral para medir la realidad— nunca podía evitar una interacción física, por minúscula que fuera, con el objeto de su observación, hasta convertirlo en parte del fenómeno estudiado y distorsionar los intentos de describirlo. La mente humana se veía obligada a elegir, entre múltiples aspectos complementarios de la realidad, cuál quería conocer con precisión en un momento dado. Una imagen completa de la realidad objetiva, si es que se disponía de ella, solo podía obtenerse combinando impresiones de aspectos complementarios de un fenómeno y teniendo en cuenta las distorsiones inherentes a cada uno de ellos.

Estas ideas revolucionarias penetraron más en la esencia de las cosas de lo que Kant o sus seguidores habían creído posible. Estamos al principio de la investigación sobre qué niveles adicionales de percepción o comprensión posibilitará la IA. Su aplicación puede permitir a los científicos colmar lagunas en la capacidad del observador humano para medir y percibir fenómenos, o en la capacidad humana (o del ordenador tradicional) para procesar grandes cantidades de datos e identificar patrones en ellos.

El mundo filosófico del siglo xx, sacudido por las disyuntivas en las fronteras de la ciencia y por la Primera Guerra Mundial, empezó a trazar nuevos caminos que se apartaban de la razón tradicional de la Ilustración y abrazaban, en cambio, la ambigüedad y la relatividad de la percepción. El filósofo austriaco Ludwig Wittgenstein, quien evitó la academia para dedicarse a la vida como jardinero y luego como maestro de escuela en un pueblo, dejó de lado la noción de

una esencia única de las cosas identificable por la razón, el objetivo perseguido por los filósofos desde Platón. En su lugar, Wittgenstein sugirió que el conocimiento se encontrara en generalizaciones sobre similitudes entre fenómenos, que él denominó «parecidos de familia»: «Y el resultado de este examen es: vemos una complicada red de semejanzas que se superponen y entrecruzan: a veces semejanzas generales, a veces semejanzas de detalle». La búsqueda de una definición y catalogación de todas las cosas, cada una con sus propios límites claramente delimitados, era errónea, sostenía. En su lugar, uno debería tratar de definir «esto y cosas similares» y familiarizarse con los conceptos resultantes, aunque tuvieran bordes «borrosos» o «indistintos».¹⁴ Más tarde, a finales del siglo xx y principios del xxi, este pensamiento sirvió de base a las teorías de la IA y el aprendizaje automático. Dichas teorías postulaban que el potencial de la IA residía en parte en su capacidad para escanear grandes conjuntos de datos para aprender tipos y patrones —por ejemplo, agrupaciones de palabras que se encuentran juntas con frecuencia, o rasgos presentes con más frecuencia en una imagen cuando tal imagen es de un gato— y luego dar sentido a la realidad identificando redes de similitudes y semejanzas con lo ya conocido por la IA. Aunque la IA nunca llegue a conocer algo de la misma forma que una mente humana, la acumulación de coincidencias con los patrones de la realidad podría aproximarse y, en ocasiones, superar el rendimiento de la percepción y la razón humanas.

El mundo de la Ilustración —con su optimismo respecto a la razón humana, a pesar de su conciencia de las trampas de la lógica humana defectuosa— ha sido durante mucho tiempo nuestro mundo. Las revoluciones científicas, sobre todo en el siglo xx, han hecho evolucionar la tecnología y la filosofía, pero ha persistido la premisa central de la Ilustración de un mundo cognoscible descubierto, paso a paso, por las mentes humanas. Hasta ahora. A lo largo de tres siglos de descubrimientos y exploraciones, los humanos han interpretado el mundo como Kant predijo que lo harían, según la estructura de sus propias mentes. Pero a medida

que el ser humano se acercaba a los límites de su capacidad cognitiva, se mostró dispuesto a recurrir a máquinas —ordenadores— para aumentar la capacidad de su pensamiento y trascender esas limitaciones. Los ordenadores añadieron un reino digital separado del reino físico en el que los humanos siempre habían vivido. Al depender cada vez más del apoyo digital, estamos entrando en una nueva época en la que la mente humana razonadora está cediendo su lugar de orgullo como única descubridora, conocedora y catalogadora de los fenómenos del mundo.

Aunque los logros tecnológicos de la era de la razón han sido significativos, hasta hace poco habían permanecido lo suficientemente esporádicos como para conciliarse con la tradición. Las innovaciones se han caracterizado como extensiones de prácticas anteriores: las películas eran fotografías en movimiento, los teléfonos eran conversaciones a través del espacio y los automóviles eran carruajes que se desplazaban rápidamente y en los que los caballos eran sustituidos por motores medidos por su «potencia». Del mismo modo, en la vida militar, los tanques eran caballería sofisticada, los aviones eran artillería avanzada, los acorazados eran fortalezas móviles y los portaaviones eran pistas de aterrizaje móviles. Incluso las armas nucleares mantuvieron la implicación de su nombre —armas— cuando las potencias nucleares organizaron sus fuerzas como artillería, para enfatizar así su experiencia previa y su comprensión de la guerra.

Pero hemos llegado a un punto de inflexión: ya no podemos concebir algunas de nuestras innovaciones como extensiones de lo ya conocido. Al comprimir el marco temporal en el cual la tecnología altera la experiencia de la vida, la revolución de la digitalización y el avance de la IA han producido fenómenos realmente nuevos, no simples versiones más potentes o eficientes de cosas pasadas. Los ordenadores, cada vez más rápidos y pequeños, se han integrado en teléfonos, relojes, servicios públicos, electrodomésticos, sistemas de seguridad, vehículos, armas e incluso cuerpos humanos. La comunicación entre estos sistemas digitales es ahora prácticamente

instantánea. Tareas que hace una generación eran manuales —leer, investigar, comprar, hablar, llevar registros, vigilar y planificar y llevar a cabo acciones militares— ahora son digitales, se basan en datos y se desarrollan en el mismo ámbito: el ciberespacio.¹⁵

Todos los niveles de la organización humana se han visto afectados por esta digitalización: a través de sus ordenadores y teléfonos, los individuos poseen (o al menos pueden acceder) a más información que nunca. Las empresas, convertidas en compiladoras de los datos de los usuarios, ejercen ahora más poder e influencia que muchos Estados soberanos. Los gobiernos, recelosos de ceder el ciberespacio a sus rivales, han entrado, explorado y empezado a explotar el reino, observando pocas reglas y ejerciendo aún menos restricciones. De hecho, se apresuran a designar el ciberespacio como un dominio en el que deben innovar para prevalecer sobre sus rivales.

Pocos han comprendido a fondo qué ha ocurrido exactamente con esta revolución digital. La velocidad tiene parte de culpa, al igual que la inundación. A pesar de sus muchos y maravillosos logros, la digitalización ha hecho que el pensamiento humano sea menos contextual y menos conceptual. Los nativos digitales no sienten la necesidad, al menos no urgente, de desarrollar conceptos que, durante la mayor parte de la historia, han compensado las limitaciones de la memoria colectiva. Pueden preguntar (y preguntan) a los motores de búsqueda lo que deseen saber, ya sea trivial, conceptual o algo intermedio. Los motores de búsqueda, a su vez, utilizan la IA para responder a sus consultas. En este proceso, los humanos delegan aspectos de su pensamiento en la tecnología. Pero la información no se explica por sí misma, sino que depende del contexto. Para que sea útil, o al menos significativa, debe entenderse a través de las lentes de la cultura y la historia.

Cuando la información se contextualiza, se convierte en conocimiento. Cuando el conocimiento impone convicciones, se convierte en sabiduría. Sin embargo, Internet inunda a los usuarios con las opiniones de miles, incluso millones, de otros usuarios,

privándoles de la soledad necesaria para esa reflexión sostenida que, históricamente, ha conducido al desarrollo de convicciones. A medida que la soledad disminuye, también lo hace la fortaleza, no solo para desarrollar convicciones, sino también para ser fiel a ellas, sobre todo cuando exigen recorrer caminos novedosos y, por tanto, a menudo solitarios. Solo las convicciones —en combinación con la sabiduría— permiten a las personas acceder a nuevos horizontes y explorarlos.

El mundo digital tiene poca paciencia para la sabiduría; sus valores están moldeados por la aprobación, no por la introspección. Desafía intrínsecamente la propuesta de la Ilustración de que la razón es el elemento más importante de la conciencia. Anulando las restricciones que históricamente han impuesto a la conducta humana la distancia, el tiempo y el lenguaje, el mundo digital propone que la conexión, en sí misma, tiene sentido.

A medida que la información en línea se ha disparado, hemos recurrido a programas informáticos que nos ayudan a clasificarla, depurarla, hacer evaluaciones basadas en patrones y guiarnos para responder a nuestras preguntas. La introducción de la IA —capaz de completar la frase que estamos enviando, identificar el libro o la tienda que buscamos e «intuir» artículos y entretenimiento que podrían gustarnos basándose en comportamientos anteriores— ha parecido a menudo más mundana que revolucionaria. Pero a medida que se va aplicando a más elementos de nuestras vidas, está alterando el papel que tradicionalmente han desempeñado nuestras mentes a la hora de dar forma, ordenar y evaluar nuestras elecciones y acciones.

¹ Edward Gibbon, *Decadencia y caída del Imperio Romano*, ed. Atalanta, Gerona, España, 2012.

² Este esfuerzo solo se consideró escandaloso en Occidente. Durante milenios, las tradiciones de gobierno y administración del Estado de otras civilizaciones han llevado a cabo investigaciones comparables sobre los intereses nacionales y los

métodos para perseguirlos: *El Arte de la Guerra*, de China, se remonta al siglo v a.C., y *El Arthashastra*, de la India, parece haber sido compuesto más o menos en la misma época.

³ El filósofo alemán de principios del siglo xx, Oswald Spengler, identificó este aspecto de la experiencia occidental de la realidad como la sociedad «fáustica», definida por su impulso al movimiento, hacia vistas expansivas del espacio y la búsqueda de un conocimiento ilimitado. Como indica el título de su principal obra, *La decadencia de Occidente*, Spengler sostenía que este impulso cultural, como todos los demás, tenía sus límites, en este caso definidos por los ciclos de la historia.

⁴ Ernst Cassirer: *La filosofía de la Ilustración*. Fondo de Cultura Económica, México, 2013.

⁵ Las tradiciones orientales llegaron antes a ideas similares, pero por caminos diferentes. El budismo, el hinduismo y el taoísmo sostenían que la experiencia de la realidad por parte de los seres humanos era subjetiva y relativa y que, por tanto, la realidad no era simplemente lo que aparecía ante sus ojos.

⁶ Baruch Spinoza: *Ética*, trad. Manuel Machado. Espuela de Plata, Sevilla, 2017.

⁷ Las vicisitudes de la historia han transformado desde entonces Königsberg en la ciudad rusa de Kaliningrado.

⁸ Immanuel Kant: *Crítica de la razón pura*, trad. Pedro Ribas. Taurus, Madrid, 2013.

⁹ Véase Paul Guyer y Allen W. Wood, *Introduction to Kant, Critique of Pure Reason*, 12.

¹⁰ Kant situó tímidamente lo divino más allá del ámbito de la razón teórica humana, preservándolo para la «creencia».

¹¹ Véase Charles Hill, *Grand Strategies: Literature, Statecraft, and World Order* (New Haven, CT: Yale University Press, 2011), 177-185.

¹² Immanuel Kant: *La paz perpetua*. Mestas ediciones, Madrid, 2007.

¹³ Michael Guillen: *Cinco ecuaciones que cambiaron el mundo: El poder y la belleza de las matemáticas*. Trad. de Francisco Páez. Editorial DEBOLSILLO, Barcelona, 2004.

¹⁴ Werner Heisenberg, «Ueber den anschaulichen Inhalt der quantentheoretischen Kinematik and Mechanik», Zeitschrift für Physik, citado en la Stanford Encyclopedia of Philosophy, «The Uncertainty Principle», <https://plato.stanford.edu/entries/qt-uncertainty/>.

¹⁵ Ludwig Wittgenstein: *Investigaciones filosóficas*, Traducción, introducción y notas de Jesús Padilla Gálvez. Editorial Trotta, Madrid, 2017.

De Turing a nuestros días

En 1943, cuando los investigadores crearon el primer ordenador moderno —electrónico, digital y programable— dieron un nuevo impulso a preguntas intrigantes: ¿Pueden pensar las máquinas? ¿Son inteligentes? ¿Pueden llegar a ser inteligentes? Estas preguntas parecían especialmente desconcertantes a la luz de los antiguos dilemas sobre la naturaleza de la inteligencia. En 1950, el matemático y descifrador de códigos Alan Turing ofreció una solución. En un artículo titulado «*Computing Machinery and Intelligence*» (Maquinaria computacional e inteligencia), Turing sugería dejar de lado el problema de la inteligencia de las máquinas. Lo importante no era el mecanismo, sino la manifestación de la inteligencia. Como la vida interior de otros seres sigue siendo incognoscible, explicaba, nuestro único medio para medir la inteligencia debería ser el comportamiento externo. Con esta idea, Turing eludió siglos de debate filosófico sobre la naturaleza de la inteligencia. El «juego de imitación» que introdujo proponía que si una máquina funcionaba tan bien que los observadores no podían distinguir su comportamiento del de un ser humano, la máquina debía considerarse inteligente.

Había nacido el test de Turing.¹

Muchos han interpretado el test de Turing literalmente, imaginando que los robots que se hacen pasar por personas (si es que eso llega a ocurrir) cumplen sus criterios. Sin embargo, cuando se aplica de forma pragmática, la prueba ha demostrado su utilidad para evaluar el rendimiento de las máquinas «inteligentes» en

actividades definidas y circunscritas, como los juegos. En lugar de exigir que sean totalmente indistinguibles de los humanos, la prueba se aplica a las máquinas cuyo rendimiento se asemeja al humano. De este modo, se centra en el rendimiento, no en el proceso. Los generadores como GPT-3 son IA porque producen un texto similar al que producen las personas, no por las características específicas de sus modelos (en el caso de GPT-3, de hecho, se entrenó utilizando grandes cantidades de información online).

En 1956, el informático John McCarthy definió, además, la inteligencia artificial como «máquinas que pueden realizar tareas características de la inteligencia humana». Las valoraciones de Turing y McCarthy sobre la IA se han convertido en puntos de referencia desde entonces, y han desplazado nuestra atención en la definición de inteligencia hacia el rendimiento (comportamiento que parece inteligente) en lugar de centrarnos en las dimensiones filosóficas, cognitivas o neurocientíficas más profundas del término.

Aunque durante el último medio siglo las máquinas han fracasado en gran medida a la hora de demostrar tal inteligencia, ese impás parece haber llegado a su fin. Tras haber funcionado durante décadas sobre la base de códigos definidos con precisión, los ordenadores producían análisis igualmente limitados por su rigidez y su naturaleza estática. Los programas tradicionales podían organizar volúmenes de datos y ejecutar cálculos complejos, pero no podían identificar imágenes de objetos sencillos ni adaptarse a entradas imprecisas. La naturaleza imprecisa y conceptual del pensamiento humano resultó ser un impedimento obstinado en el desarrollo de la IA. En la última década, sin embargo, las innovaciones informáticas han creado IA que han empezado a igualar o superar los logros humanos en esos campos.

Las IA son imprecisas, dinámicas, emergentes y capaces de «aprender». Las IA «aprenden» consumiendo datos y extrayendo observaciones y conclusiones a partir de ellos. Mientras que los sistemas anteriores requerían entradas y salidas exactas, las IA con funciones imprecisas no necesitan ni lo uno ni lo otro. Estas IA

traducen textos no intercambiando palabras sueltas, sino identificando y empleando frases y patrones idiomáticos. Asimismo, estas IA se consideran dinámicas porque evolucionan en respuesta a circunstancias cambiantes y emergentes y porque pueden identificar soluciones novedosas para los humanos. En el campo de las máquinas, estas cuatro cualidades son revolucionarias.

Pensemos, por ejemplo, en la irrupción de AlphaZero en el mundo del ajedrez. Los programas de ajedrez clásicos se basaban en la experiencia humana, desarrollada por el juego humano y codificada en su programación. Pero AlphaZero desarrolló sus habilidades jugando millones de partidas contra sí mismo, a partir de las cuales descubrió patrones por sí mismo.

Los componentes básicos de estas técnicas de «aprendizaje» son los algoritmos, conjuntos de pasos que traducen datos de entrada (las reglas de un juego o medidas de la calidad de las jugadas dentro de esas reglas) en resultados repetibles (ganar la partida). Pero los algoritmos de aprendizaje automático se alejan de la precisión y previsibilidad de los algoritmos clásicos, incluidos los de cálculos como la división larga. A diferencia de los algoritmos clásicos, que consisten en pasos para producir resultados precisos, los algoritmos de aprendizaje automático se basan en pasos para mejorar los resultados imprecisos. Estas técnicas están progresando notablemente.

La aviación es otro ejemplo. Pronto, la IA pilotará o copilotará diversos vehículos en el aire. En el programa *AlphaDogfight*, de DARPA, los pilotos de cazas de IA han superado a los humanos en combates simulados, ejecutando maniobras fuera del alcance de las capacidades de los pilotos humanos. Ya sea pilotando aviones para combatir en guerras o drones para repartir comida, la IA está llamada a tener un impacto significativo en el futuro de la aviación militar y civil.

Aunque solo hemos visto los comienzos de estas innovaciones, ya han alterado sutilmente el tejido de la experiencia humana. Y en las próximas décadas, la tendencia no hará sino acelerarse.

Como los conceptos tecnológicos que impulsan la transformación de la IA son tan complejos e importantes, este capítulo se dedicará a explicar tanto la evolución como el estado actual de los distintos tipos de aprendizaje y uso de las máquinas, tanto asombrosamente potentes como intrínsecamente limitados. Una introducción básica a su estructura, capacidades y limitaciones es vital para comprender los cambios sociales, culturales y políticos que ya han traído consigo, así como los cambios que probablemente producirán en el futuro.

La evolución de la IA

La humanidad siempre ha soñado con un asistente, una máquina capaz de realizar tareas con la misma competencia que un humano. En la mitología griega, el divino herrero Hefesto forjó robots capaces de realizar tareas humanas, como el gigante de bronce Talos, que patrullaba las costas de Creta y la protegía de las invasiones. Luis XIV de Francia, en el siglo XVII, y Federico el Grande, de Prusia, en el XVIII, sintieron fascinación por los autómatas mecánicos y supervisaron la construcción de prototipos. En realidad, sin embargo, diseñar una máquina y hacerla capaz de una actividad útil —incluso con la llegada de la informática moderna— ha resultado endiabladamente difícil. Uno de los principales retos es cómo —y qué— enseñarle.

Los primeros intentos de crear inteligencias artificiales útiles en la práctica codificaban explícitamente los conocimientos humanos (mediante conjuntos de reglas o hechos) en sistemas informáticos. Pero gran parte del mundo no está organizado de forma discreta ni es fácilmente reducible a reglas simples o representaciones simbólicas. Mientras en los campos que sí utilizan una caracterización precisa —el ajedrez, la manipulación algebraica y la automatización de procesos empresariales— la IA experimentó grandes avances, en otros, como la traducción de idiomas y el reconocimiento visual de objetos, la ambigüedad inherente paralizó el progreso.

Los retos del reconocimiento visual de objetos ilustran las deficiencias de estos primeros programas. Incluso los niños pequeños pueden identificar imágenes con facilidad. Pero las primeras generaciones de IA no podían. Al principio, los programadores intentaron destilar las características distintivas de un objeto en una representación simbólica. Por ejemplo, para enseñar a la IA a identificar la imagen de un gato, los programadores crearon representaciones abstractas de los distintos atributos (bigotes, orejas puntiagudas, cuatro patas, un cuerpo) de un gato idealizado. Pero los gatos distan mucho de ser estáticos: pueden enroscarse, correr y estirarse y poseer diversos tamaños y colores. En la práctica, la formulación de modelos abstractos y su posterior adaptación a datos muy variables ha resultado prácticamente inviable.

Como estos sistemas formalistas e inflexibles solo tenían éxito en ámbitos cuyas tareas podían realizarse mediante la codificación de reglas claras, desde finales de los 80 hasta los 90, el campo entró en un periodo denominado «invierno de la IA». Aplicada a tareas más dinámicas, la IA resultó frágil y produjo resultados que no superaban la prueba de Turing, es decir, que no alcanzaban o imitaban el rendimiento humano. Como las aplicaciones de estos sistemas eran limitadas, la financiación de I+D disminuyó y los avances se ralentizaron.

Entonces, en la década de 1990, se produjo un gran avance. En el fondo, la IA consiste en realizar tareas, en crear máquinas capaces de idear y ejecutar soluciones competentes a problemas complejos. Los investigadores se percataron de que era necesario un nuevo enfoque que permitiera a las máquinas aprender por sí mismas. En pocas palabras, se produjo un cambio conceptual: pasamos de intentar codificar en las máquinas conocimientos destilados por el ser humano a delegar en ellas el propio proceso de aprendizaje.

En la década de 1990, un grupo de investigadores renegados dejó de lado muchos de los supuestos de la era anterior y se centró en el aprendizaje automático. Aunque el aprendizaje automático se

remontaba a los años 50, los nuevos avances permitieron aplicaciones prácticas. Los métodos que mejor han funcionado en la práctica extraen patrones de grandes conjuntos de datos mediante redes neuronales. En términos filosóficos, los pioneros de la IA habían abandonado el enfoque de los primeros ilustrados de reducir el mundo a reglas mecanicistas para construir aproximaciones a la realidad. Para identificar la imagen de un gato, se dieron cuenta de que una máquina tenía que «aprender» una serie de representaciones visuales de gatos observando al animal en diversos contextos. Para permitir el aprendizaje automático, lo que importaba era el solapamiento entre las diversas representaciones de una cosa, no su ideal: en términos filosóficos, Wittgenstein, en vez de Platón. Así nació el moderno campo del aprendizaje automático, el de los programas que aprenden a través de la experiencia.

La IA moderna

A continuación se produjeron avances significativos. En la década del 2000, en el campo del reconocimiento visual de objetos, cuando los programadores desarrollaron IA para representar una aproximación de un objeto mediante el aprendizaje a partir de un conjunto de imágenes —algunas de las cuales contenían el objeto, otras no—, las IA identificaron los objetos con mucha más eficacia que sus predecesoras codificadas.

La IA utilizada para identificar la halicina ilustra la importancia del proceso de aprendizaje automático. Cuando los investigadores del MIT diseñaron un algoritmo de aprendizaje automático para predecir las propiedades antibacterianas de las moléculas, entrenando el algoritmo con un conjunto de datos de más de dos mil moléculas, el resultado fue algo que ningún algoritmo convencional —y ningún ser humano— podría haber logrado. Los humanos no solo no entienden las numerosas conexiones que la IA revela entre las propiedades de un compuesto y su capacidad antibiótica, sino que, algo aún más importante, las propiedades en sí no pueden expresarse en forma de reglas. Sin embargo, un algoritmo de aprendizaje automático que

mejora un modelo basándose en los datos subyacentes es capaz de reconocer las relaciones que los humanos han pasado por alto.

Como se ha señalado anteriormente, dicha IA es imprecisa en el sentido de que no requiere una relación predefinida entre una propiedad y un efecto para identificar una relación parcial. Puede, por ejemplo, seleccionar candidatos muy probables de entre un conjunto más amplio de posibles candidatos. Esta capacidad representa uno de los elementos vitales de la IA moderna. Al utilizar el aprendizaje automático para crear y ajustar modelos basados en información del mundo real, la IA moderna puede aproximar resultados y analizar ambigüedades que habrían bloqueado a los algoritmos clásicos. Al igual que un algoritmo clásico, un algoritmo de aprendizaje automático consta de una secuencia de pasos precisos. Pero esos pasos no producen directamente un resultado específico, como en un algoritmo clásico. Por el contrario, los algoritmos modernos de IA miden la calidad de los resultados y proporcionan medios para mejorarlos, lo cual permite aprenderlos en lugar de especificarlos directamente.

Las redes neuronales, inspiradas en la estructura del cerebro humano (aunque, debido a su complejidad, no siguen su modelo por completo), impulsan la mayoría de estos avances. En 1958, Frank Rosenblatt, investigador del Laboratorio Aeronáutico de Cornell, tuvo una idea: ¿Podrían los científicos desarrollar un método de codificación de la información similar al del cerebro humano, que codifica la información conectando aproximadamente cien mil millones de neuronas con cuatrillones — 10^{12} — de sinapsis? Decidió intentarlo: diseñó una red neuronal artificial que codificaba relaciones entre nodos (análogos a las neuronas) y pesos numéricos (análogos a las sinapsis). Se trata de redes que codifican la información mediante una estructura de nodos —y las conexiones entre esos nodos— en la cual los pesos designados representan la fuerza de las conexiones entre nodos. Durante décadas, la falta de potencia de cálculo y de algoritmos sofisticados frenó el desarrollo de todas las redes neuronales, salvo las rudimentarias. Sin embargo,

los avances en ambos campos han liberado a los desarrolladores de IA de estas restricciones.

En el caso de la halicina, una red neuronal captó la asociación entre las moléculas (las entradas) y su potencial para inhibir el crecimiento bacteriano (la salida). La IA que descubrió la halicina lo hizo sin información sobre los procesos químicos o las funciones de los fármacos, descubriendo relaciones entre las entradas y las salidas a través del aprendizaje profundo, en el cual las capas de una red neuronal más cercanas a la entrada tienden a reflejar aspectos de la entrada, mientras las capas más alejadas de la entrada tienden a reflejar generalizaciones más amplias, predictivas de la salida deseada.

El aprendizaje profundo permite a las redes neuronales captar relaciones complejas como las existentes entre la eficacia de los antibióticos y aspectos de la estructura molecular reflejados en los datos de entrenamiento (peso atómico, composición química, tipos de enlaces y similares). Esta red permite a la IA captar conexiones intrincadas, incluso aquellas que se le han escapado a los humanos. En su fase de entrenamiento, a medida que la IA recibe nuevos datos, ajusta los pesos en toda la red. La precisión de la red depende tanto del volumen como de la calidad de los datos con los que se entrena. A medida que la red recibe más datos y se compone de más capas de red, los pesos empiezan a captar las relaciones con mayor precisión. Las redes profundas actuales suelen contener alrededor de diez capas.

Pero el entrenamiento de redes neuronales consume muchos recursos. El proceso requiere una potencia de cálculo considerable y algoritmos complejos para analizar y ajustarse a grandes cantidades de datos. A diferencia de los humanos, la mayoría de las IA no pueden entrenar y ejecutar simultáneamente. Más bien dividen su esfuerzo en dos pasos: entrenamiento e inferencia. Durante la fase de entrenamiento, los algoritmos de medición y mejora de la calidad de la IA evalúan y modifican su modelo para obtener resultados de calidad. En el caso de la halicina, en dicha fase la IA identifica las

relaciones entre las estructuras moleculares y los efectos antibióticos a partir de los datos del conjunto de entrenamiento. Después, en la fase de inferencia, los investigadores encargaron a la IA que identificara los antibióticos que su modelo recién entrenado predecía que tendrían un fuerte efecto antibiótico. La IA no llegó a conclusiones razonando como los humanos, sino aplicando el modelo que había desarrollado.

Diferentes tareas, diferentes estilos de aprendizaje

Como la aplicación de la IA varía en función de las tareas que realiza, también deben variar las técnicas de los desarrolladores para crear esa IA. Este es un reto fundamental de la implementación del aprendizaje automático: diferentes objetivos y funciones requieren diferentes técnicas de formación. Pero de la combinación de métodos —algoritmos de aprendizaje automático, redes neuronales y técnicas de aprendizaje— surgen nuevas posibilidades, como, por ejemplo, las IA que detectan el cáncer.

En el momento de escribir estas líneas, destacan tres formas de aprendizaje automático: el aprendizaje supervisado, el aprendizaje no supervisado y el aprendizaje por refuerzo. El aprendizaje supervisado produjo la IA que descubrió la halicina. Para recapitular, cuando los investigadores del MIT quisieron identificar nuevos antibióticos potenciales, utilizaron una base de datos de dos mil moléculas para entrenar un modelo en el que la estructura molecular era la entrada y la eficacia antibiótica era la salida. Los investigadores presentaron a la IA las estructuras moleculares, cada una etiquetada según su eficacia antibiótica. A continuación, a partir de nuevos compuestos, la IA estimó la eficacia antibiótica.

Esta técnica se denomina aprendizaje supervisado porque los desarrolladores de IA utilizaron un conjunto de datos que contenía entradas de ejemplo (en este caso, estructuras moleculares) etiquetadas individualmente según la salida o resultado deseado (en este caso, eficacia como antibiótico). Los desarrolladores han utilizado técnicas de aprendizaje supervisado para muchos fines,

como la creación de IA que reconocen imágenes. Para esta tarea, las IA se entrenan con un conjunto de imágenes preetiquetadas, aprendiendo a asociar una imagen con su etiqueta apropiada; por ejemplo, la imagen de un gato con la etiqueta «gato». Una vez codificada la relación entre imágenes y etiquetas, las IA son capaces de identificar correctamente nuevas imágenes. Por lo tanto, cuando los desarrolladores disponen de un conjunto de datos que indica un resultado deseado para cada conjunto de entradas, el aprendizaje supervisado ha demostrado ser un modo especialmente eficaz de crear un modelo capaz de predecir resultados en respuesta a nuevas entradas.

Sin embargo, en situaciones en las que los desarrolladores solo disponen de una gran cantidad de datos, pueden emplear el aprendizaje no supervisado para extraer información potencialmente útil. Gracias a Internet y a la digitalización de la información, las empresas, los gobiernos y los investigadores están inundados de datos, a los que pueden acceder más fácilmente que en el pasado. Los profesionales del marketing tienen más información sobre sus clientes, los biólogos más datos sobre el ADN y los banqueros más transacciones financieras archivadas. Cuando los profesionales del marketing quieren identificar su base de clientes, o cuando los analistas de fraudes buscan posibles incoherencias entre montones de transacciones, el aprendizaje no supervisado permite a las IA identificar patrones o anomalías sin tener ninguna información sobre los resultados. En el aprendizaje no supervisado, los datos de entrenamiento solo contienen datos de entrada. A continuación, los programadores encargan al algoritmo de aprendizaje que produzca agrupaciones basadas en algún peso específico para medir el grado de similitud. Por ejemplo, los servicios de *streaming* de vídeo como Netflix utilizan algoritmos para identificar grupos de clientes con hábitos de visionado similares con el fin de recomendarles otros servicios. Pero afinar esos algoritmos puede ser complejo: como la mayoría de las personas tienen varios intereses, suelen agruparse en varios clústeres.

Las IA entrenadas mediante aprendizaje no supervisado pueden identificar patrones que los humanos pasarían por alto debido a la sutileza del patrón, a la escala de los datos o a ambas cosas. Como estas IA se entrenan sin especificar los resultados «adecuados», pueden —al igual que un humano autodidacta— producir ideas sorprendentemente innovadoras. Sin embargo, tanto el autodidacta humano como estas IA pueden producir resultados excéntricos y sin sentido.

Tanto en el aprendizaje no supervisado como en el supervisado, las IA utilizan principalmente datos para realizar tareas como descubrir tendencias, identificar imágenes y hacer predicciones. Más allá del análisis de datos, los investigadores trataron de entrenar a las IA para operar en entornos dinámicos. Así nació una tercera gran categoría de aprendizaje automático, el aprendizaje por refuerzo.

En el aprendizaje por refuerzo, la IA no es pasiva e identifica las relaciones entre los datos. En cambio, la IA es un «agente» en un entorno controlado, que observa y registra las respuestas a sus acciones. Por lo general, se trata de versiones simuladas y simplificadas de la realidad que carecen de las complejidades del mundo real. Es más fácil simular con precisión el funcionamiento de un robot en una cadena de montaje que el caos de una calle abarrotada. Pero incluso en un entorno simulado y simplificado, como una partida de ajedrez, un solo movimiento puede desencadenar una cascada de oportunidades y riesgos. En consecuencia, dirigir una IA para que se entrene en un entorno artificial es, en general, insuficiente para obtener el mejor rendimiento. Se necesita retroalimentación.

Proporcionar esa retroalimentación es la tarea de la función de recompensa, que indica a la IA el éxito de su enfoque. Ningún ser humano podría desempeñar eficazmente esta función: al funcionar con procesadores digitales, las IA pueden entrenarse cientos, miles o miles de millones de veces en el espacio de horas o días, lo cual hace que la retroalimentación humana directa sea totalmente impracticable. En su lugar, los programadores automatizan estas

funciones de recompensa, especificando cuidadosamente cómo opera la función y la naturaleza de cómo simula la realidad. Lo ideal es que el simulador proporcione una experiencia realista y que la función de recompensa promueva decisiones eficaces.

El simulador de AlphaZero era sencillo: jugaba contra sí mismo. Luego, para evaluar su rendimiento, empleaba una función de recompensa³ que puntuaba sus movimientos en función de las oportunidades que creaban. El aprendizaje por refuerzo requiere la participación humana en la creación del entorno de entrenamiento de la IA (aunque no proporcione información directa durante el propio entrenamiento): los humanos definen un simulador y una función de recompensa, y la IA se entrena a sí misma sobre esa base. Para obtener resultados significativos, es vital especificar cuidadosamente el simulador y la función de recompensa.

El poder del aprendizaje automático

A partir de estos pocos componentes básicos surgen innumerables aplicaciones. En agricultura, la IA facilita la administración precisa de pesticidas, la detección de enfermedades y la predicción del rendimiento de las cosechas. En medicina, facilita el descubrimiento de nuevos fármacos, la identificación de nuevas aplicaciones de los fármacos existentes y la detección o predicción de enfermedades futuras. En el momento de escribir estas líneas, la IA ha detectado el cáncer de mama antes que los médicos humanos, identificando sutiles indicadores radiológicos; ha detectado la retinopatía, una de las principales causas de ceguera, analizando fotos de la retina; ha predicho la hipoglucemia en diabéticos, analizando historiales médicos; y ha detectado otras afecciones hereditarias analizando códigos genéticos. En finanzas, la IA está equipada para facilitar procesos de gran volumen: aprobación (o denegación) de préstamos, adquisiciones, fusiones, declaraciones de quiebra y otras transacciones.

En otros campos, facilita la transcripción y la traducción: en cierto modo, la ilustración más convincente de todas. Durante milenios, la humanidad se ha enfrentado a la incapacidad de las personas para comunicarse con claridad a través de las barreras culturales y lingüísticas. La incomprensión mutua y la inaccesibilidad de la información en una lengua para el hablante de otra han provocado malentendidos, obstaculizado el comercio y fomentado la guerra. La historia de la Torre de Babel es un símbolo de la imperfección humana y un amargo castigo por la arrogancia humana. Ahora, parece que la IA está a punto de poner a disposición de un amplio público potentes capacidades de traducción, y permitir así que muchas más personas se comuniquen más fácilmente.

Hasta los años 90, los investigadores intentaron crear programas de traducción de idiomas basados en reglas. Aunque sus esfuerzos tuvieron cierto éxito en el laboratorio, no dieron buenos resultados en el mundo real. La variabilidad y sutileza del lenguaje no se reducían a reglas sencillas. Todo esto cambió cuando, en 2015, los desarrolladores empezaron a aplicar redes neuronales profundas al problema. De repente, la traducción automática dio un salto adelante. Pero su mejora no solo se derivó de la aplicación de redes neuronales o técnicas de aprendizaje automático. Más bien surgió de aplicaciones nuevas y creativas de estos enfoques. Estos avances subrayan un punto clave: a partir de los componentes básicos del aprendizaje automático, los desarrolladores tienen la capacidad de seguir innovando de forma brillante, desbloqueando nuevas IA en el proceso.

Para traducir un idioma a otro, un traductor necesita captar patrones específicos: dependencias secuenciales. Las redes neuronales estándar discernen patrones de asociación entre entradas y salidas, como los conjuntos de propiedades químicas que suelen poseer los antibióticos. Pero esas redes no captan, sin modificaciones, las dependencias secuenciales, como la probabilidad de que una palabra aparezca en una posición determinada de una frase teniendo en cuenta las palabras que la preceden. Por ejemplo,

si una frase empieza con las palabras «Fui a pasear al», es mucho más probable que la siguiente palabra sea «perro» que «gato» o «avión». Para captar estas dependencias secuenciales, los investigadores idearon redes que utilizan como entradas no solo el texto aún por traducir, sino también el texto que ya ha sido traducido. De este modo, la IA puede identificar la siguiente palabra basándose en dependencias secuenciales en el idioma de entrada y en el idioma al que se está traduciendo el texto. Las más potentes de estas redes son las *transformadoras*, que no necesitan procesar el lenguaje de izquierda a derecha. El BERT de Google, por ejemplo, es un transformador bidireccional diseñado para mejorar las búsquedas.

Además, en un cambio considerable con respecto al aprendizaje supervisado convencional, los investigadores de la traducción de idiomas emplearon «corpus paralelos», una técnica en la cual la correspondencia específica entre entradas y salidas (por ejemplo, el significado entre textos en dos o más idiomas) no es necesaria para el entrenamiento. En los enfoques convencionales, los desarrolladores entrenaban la IA utilizando textos y sus traducciones preexistentes; al fin y al cabo, tenían el nivel necesario de correspondencia entre una lengua y otra. Sin embargo, este enfoque limitaba enormemente la cantidad de datos de entrenamiento, así como los tipos de texto disponibles: aunque los textos gubernamentales y los libros más vendidos son traducidos con frecuencia, las publicaciones periódicas, las redes sociales, los sitios web y otros escritos informales generalmente no lo son.

En lugar de limitar el entrenamiento de las IA a textos cuidadosamente traducidos, los investigadores simplemente suministran artículos y otros textos en varios idiomas sobre un mismo tema, sin preocuparse de hacer traducciones detalladas entre ellos. La técnica de los corpus paralelos consiste en entrenar a las IA con corpus de textos que coincidan aproximadamente, pero sin traducir. Es como pasar de una clase introductoria de idiomas a un programa de inmersión total. El entrenamiento es menos preciso,

pero el volumen de datos disponibles es mucho mayor: los desarrolladores pueden incluir artículos de prensa, reseñas de libros y películas, relatos de viajes y prácticamente cualquier otra publicación formal o informal sobre un tema tratado por escritores en muchos idiomas. El éxito de este enfoque ha conducido a un uso más generalizado del aprendizaje parcialmente supervisado, en el cual se utiliza información muy aproximada o parcial para el entrenamiento.

Cuando *Google Translate* empezó a emplear redes neuronales profundas entrenadas con corpus paralelos, su rendimiento mejoró un 60 %, y ha seguido mejorando desde entonces.

El avance radical de la traducción automática de idiomas promete transformar los negocios, la diplomacia, los medios de comunicación, el mundo académico y otros campos, porque las personas se relacionan con lenguas diferentes a las suyas de forma más fácil, rápida y barata que nunca.

Por supuesto, la capacidad de traducir textos y clasificar imágenes es una cosa. La capacidad de generar —crear— nuevos textos, imágenes y sonidos es otra cosa. Hasta ahora, las IA que hemos descrito destacan en la identificación de soluciones: una victoria en ajedrez, un candidato a fármaco, una traducción lo suficientemente buena como para usarla. Pero otra técnica, las redes neuronales generativas, es capaz de crear. Primero, las redes neuronales generativas se entrenan utilizando texto o imágenes. A continuación, producen nuevos textos o imágenes, sintéticos pero realistas. Por ejemplo: una red neuronal estándar puede identificar una imagen de un rostro humano, pero una red *generativa* puede crear una imagen de un rostro humano que *parezca* real. Conceptualmente, se apartan de sus predecesoras.

Las aplicaciones de los llamados generadores son asombrosas. Si se aplican con éxito a la codificación o la escritura, un autor podría simplemente crear un esquema y dejar que el generador rellene los detalles. O un publicista o cineasta podría suministrar a un generador unas cuantas imágenes o un guion gráfico, y luego dejar

que la IA creara un anuncio o comercial sintético. Y lo que es más preocupante, los generadores también podrían utilizarse para crear falsificaciones profundas: representaciones falsas, indistinguibles de la realidad, de personas que hacen o dicen cosas que nunca han hecho o dicho. Los generadores enriquecerán nuestro espacio informativo, pero sin controles, es probable que también difuminen la línea que separa la realidad de la fantasía.

Una técnica de entrenamiento habitual para la creación de IA generativa enfrenta a dos redes con objetivos de aprendizaje complementarios. Estas redes se denominan redes generativas adversarias o GAN (por sus siglas en inglés). El objetivo de la red *generadora* es crear resultados potenciales, mientras que el objetivo de la red *discriminadora* es evitar que se generen resultados deficientes. Por analogía, se puede pensar que el generador se encarga de la lluvia de ideas y el discriminador de evaluar qué ideas son relevantes y realistas. En la fase de entrenamiento, el generador y el discriminador se entrenan alternativamente, manteniendo fijo el generador para entrenar al discriminador y viceversa.

Estas técnicas no son perfectas —entrenar GAN puede ser un reto y, a menudo, generar malos resultados—, pero las IA que producen pueden lograr hazañas notables. En su forma más común, las IA entrenadas con GAN pueden sugerir frases completas al redactar correos electrónicos o permitir que los motores de búsqueda completen consultas parciales. Y lo que es más espectacular, las GAN pueden utilizarse para desarrollar IA capaces de completar los detalles de un código esbozado. En otras palabras, los programadores pronto podrán esbozar el programa deseado y entregarlo a una IA para que lo complete.

En la actualidad, GPT-3, capaz de producir textos similares a los humanos (véase el capítulo 1), es una de las IA generativas más destacadas. Extiende a la producción lingüística el enfoque que transformó la traducción de idiomas. Dadas unas pocas palabras, puede «extrapolar» para producir una frase, o dada una frase temática, puede extrapolarse para producir un párrafo. Los

transformadores como GPT-3 detectan patrones en elementos secuenciales como el texto, lo cual les permite predecir y generar los elementos que probablemente seguirán. En el caso de GPT-3, la IA puede captar las dependencias secuenciales entre palabras, párrafos o código para generar estos resultados.

Entrenados con grandes cantidades de datos extraídos principalmente de Internet, los transformadores también pueden transformar texto en imágenes y viceversa, ampliar y condensar descripciones y realizar tareas similares. Hoy en día, la calidad de los resultados de GPT-3 —y de otras IA similares— puede ser impresionante, pero muy variable. A veces, sus resultados parecen muy inteligentes; otras, tontos o incluso completamente ininteligibles. Sin embargo, la función básica de los transformadores puede alterar muchos campos, incluidos los creativos. Por eso despiertan tanto interés entre investigadores y desarrolladores, quienes estudian sus puntos fuertes, sus limitaciones y sus aplicaciones.

El aprendizaje automático no solo ha ampliado la aplicabilidad de la IA, sino que la ha revolucionado incluso en ámbitos en los cuales los enfoques anteriores, como los sistemas simbólicos y basados en reglas, tenían éxito. Los métodos de aprendizaje automático han llevado a la IA de derrotar a expertos humanos en ajedrez a descubrir estrategias de ajedrez totalmente nuevas. Y su capacidad de descubrimiento no se limita a los juegos. Como hemos mencionado, DeepMind construyó una IA que consiguió reducir el gasto energético de los centros de datos de Google en un 40 % más de lo que sus excelentes ingenieros podían lograr. Este y otros avances están llevando a la IA más allá de lo que Turing imaginó en su prueba —un rendimiento indistinguible de la inteligencia humana— para alcanzar un rendimiento que supera a los humanos, y ampliar así las fronteras de la comprensión. Tales avances prometen permitir que la IA se encargue de nuevas tareas, que se generalice e incluso que genere textos y códigos originales.

Por supuesto, siempre que una tecnología se hace más potente o prevalente, los retos acompañan a estos avances. La personalización de la búsqueda —la función en línea que la mayoría de nosotros empleamos con más frecuencia— ilustra muy bien esto. En el capítulo 1, describimos la diferencia entre una búsqueda tradicional en Internet y una búsqueda gestionada por IA como la diferencia entre estar expuesto a ropa de diseño y estar expuesto a toda la gama de ropa disponible para la compra. Una IA permite este resultado —un motor de búsqueda que se adapta a un usuario individual— de dos maneras: (1) tras recibir consultas, como «cosas que hacer en Nueva York», una IA puede producir conceptos, como «pasear por Central Park» y «ver un espectáculo en Broadway», y (2) la IA puede recordar las cosas que se le han preguntado antes a un motor de búsqueda y los conceptos que, en respuesta, ha producido. Además, puede almacenar estos conceptos en su versión de la memoria. Con el tiempo, puede utilizar su memoria para producir conceptos cada vez más específicos —y, en teoría, cada vez más útiles— para sus usuarios. Los servicios de *streaming* en línea hacen lo mismo, utilizando la IA para hacer sugerencias de programas de televisión y películas «más» (más centradas, más positivas, o más cualquier cosa que la gente quiera que sean). Esto puede ser beneficioso. La IA puede alejar a los niños de contenidos adultos y, al mismo tiempo, dirigirlos hacia contenidos apropiados para su edad o marco de referencia. La IA puede alejarnos a todos de contenidos violentos, explícitos u ofensivos para nuestra sensibilidad. Depende de lo que los algoritmos, tras analizar las acciones pasadas de los usuarios, deduzcan que son sus preferencias. A medida que la IA va conociendo a las personas, el resultado es en gran medida positivo: los abonados a los servicios de *streaming*, por ejemplo, tienen cada vez más probabilidades de ver programas y películas que les interesan, en lugar de ofenderles o confundirles.

La propuesta de que ese filtro puede ayudar a orientar las elecciones es familiar y práctica. En el mundo físico, los turistas en

países extranjeros pueden contratar guías que les muestren los lugares más históricos o los más significativos según su religión, nacionalidad o profesión. Pero tal filtro puede convertirse en censura por omisión. Un guía puede evitar los barrios marginales y las zonas de alta criminalidad. En un país autoritario, un guía puede ser un «guardián del gobierno» y mostrar al turista solo lo que el régimen quiere que vea. En el ciberespacio, la filtración se refuerza a sí misma. Cuando la lógica algorítmica que personaliza la búsqueda y el *streaming* empieza a personalizar el consumo de noticias, libros u otras fuentes de información, amplifica algunos temas y fuentes de información y, por necesidad práctica, omite otros por completo. La consecuencia de la omisión de facto es doble: puede crear cámaras de resonancia personales y puede fomentar la discordancia entre ellas. Lo que una persona consume (y, por tanto, supone que refleja la realidad) difiere de lo que consume una segunda persona, y lo que consume una segunda persona difiere aún más de lo que consume una tercera, una paradoja que analizaremos en el capítulo 6.

Gestionar los riesgos que planteará una IA cada vez más extendida es una tarea que debe llevarse a cabo paralelamente al avance de este campo, y es una de las razones de este libro. Todos debemos prestar atención a los riesgos potenciales de la IA. No podemos dejar su desarrollo o aplicación en manos de un único grupo, ya sean investigadores, empresas, gobiernos u organizaciones de la sociedad civil.

Límites y gestión de la IA

A diferencia de las generaciones anteriores de IA, en las que las personas resumían la comprensión de la realidad de una sociedad en el código de un programa, las IA contemporáneas de aprendizaje automático modelan en gran medida la realidad por sí solas. Aunque los desarrolladores pueden examinar los resultados generados por sus IA, estas no «explican» cómo o qué han aprendido en términos humanos. Los desarrolladores tampoco pueden pedir a una IA que

describa lo que ha aprendido. Al igual que ocurre con los humanos, no se puede saber realmente qué se ha aprendido y por qué (aunque los humanos a menudo pueden ofrecer algunas explicaciones o justificaciones; en el momento de escribir estas líneas, las IA no pueden). En el mejor de los casos, solo podemos observar los resultados que produce una IA una vez que ha completado su entrenamiento. Por consiguiente, los humanos deben trabajar hacia atrás. Una vez que una IA produce un resultado, las personas —ya sean investigadores o auditores— deben verificar que la IA produce los resultados deseados.

A veces, al operar más allá de los límites de la experiencia humana y ser incapaces de conceptualizar o generar explicaciones, las IA pueden producir descubrimientos que son ciertos pero que están más allá de las fronteras de la comprensión humana (al menos de la actual). Cuando las IA producen descubrimientos inesperados de este tipo, los humanos pueden encontrarse en una situación similar a la de Alexander Fleming, el descubridor de la penicilina. En el laboratorio de Fleming, un moho productor de penicilina colonizó accidentalmente una placa de Petri, eliminando las bacterias causantes de enfermedades y descubriendo a Fleming la existencia de un potente compuesto desconocido hasta entonces. En aquel momento, la humanidad, que carecía del concepto de antibiótico, no comprendía cómo funcionaba la penicilina. El descubrimiento abrió todo un campo de investigación. Las IA producen descubrimientos igual de sorprendentes —como la identificación de candidatos a fármacos y nuevas estrategias para ganar partidos—, dejando en manos de los humanos la tarea de adivinar su significado y, si es prudente, integrar estos descubrimientos en los corpus de conocimiento existentes.

Además, la IA no puede reflexionar sobre lo que descubre. A lo largo de muchas épocas, los seres humanos han vivido la guerra y luego han reflexionado sobre sus lecciones, sus penas y sus extremos: desde el relato de Homero sobre Héctor y Aquiles a las puertas de Troya en La Ilíada, hasta el retrato de Picasso de las

víctimas civiles de la Guerra Civil española en el Guernica. La IA no puede hacerlo, ni siente la obligación moral o filosófica de hacerlo. Simplemente aplica su método y produce un resultado, sea este — desde una perspectiva humana— banal o impactante, benigno o maligno. La IA no puede reflexionar; la importancia de sus acciones debe decidirla el ser humano. Por lo tanto, los humanos deben regular y supervisar la tecnología.

La incapacidad de la IA para contextualizar o reflexionar como un ser humano hace que sea especialmente importante prestar atención a sus retos. El software de reconocimiento de imágenes de Google ha etiquetado erróneamente imágenes de personas como animales⁴ y de animales como armas⁵. Las IA no solo son incapaces de reflexionar, sino que también cometen errores, incluso errores que cualquier ser humano consideraría rudimentarios. Y aunque los desarrolladores no cesan de eliminar fallos, a menudo la implantación ha precedido a la resolución de problemas.

Estos errores de identificación tienen varias causas. Uno de ellos es el sesgo del conjunto de datos. El aprendizaje automático requiere datos, sin los cuales las IA no pueden aprender buenos modelos. Un problema crítico es que, si no se presta la debida atención, muy probablemente se produzcan problemas de insuficiencia de datos en el caso de grupos infrarrepresentados, como las minorías raciales. En concreto, los sistemas de reconocimiento facial se han entrenado a menudo en conjuntos de datos con un número desproporcionadamente bajo de imágenes de personas de raza negra, lo cual ha dado lugar a una escasa precisión. Tanto la cantidad como la cobertura son importantes: entrenar a las IA con grandes cantidades de imágenes muy similares dará lugar a redes neuronales que no estén seguras de un resultado porque no se lo han encontrado antes. En otras situaciones de alto riesgo, puede ocurrir algo similar. Por ejemplo, los conjuntos de datos para el entrenamiento de coches autónomos pueden contener relativamente pocos ejemplos de situaciones extraordinarias, como cuando un ciervo salta sobre la carretera, dejando a la IA sin

especificar cómo enfrentarse a ese escenario. Sin embargo, en tales situaciones, la IA tiene que funcionar al máximo nivel.

Otra posibilidad es que el sesgo de la IA se derive directamente del sesgo humano, es decir, que sus datos de entrenamiento contengan sesgos inherentes a las acciones humanas. Esto puede ocurrir al etiquetar los resultados del aprendizaje supervisado: cualquier error de identificación que cometa el etiquetador, deliberado o involuntario, será codificado por la IA. O un desarrollador puede especificar incorrectamente una función de recompensa utilizada en el entrenamiento por refuerzo. Imaginemos una IA entrenada para jugar al ajedrez en un simulador que sobrevalora un conjunto de jugadas preferidas por su creador. Al igual que su creador, la IA aprenderá a preferir esas jugadas, aunque en la práctica le salgan mal.

Por supuesto, el problema del sesgo en la tecnología no se limita a la IA. El pulsioxímetro, que se ha convertido en una medida cada vez más pertinente de dos parámetros de salud —la frecuencia cardíaca y la saturación de oxígeno— desde el inicio de la pandemia de COVID-19, sobrestima la saturación de oxígeno en las personas de piel oscura. Al asumir que la forma en que la piel clara absorbe la luz es «normal», sus diseñadores asumieron efectivamente que la forma en que la piel oscura absorbe la luz es «anormal». El pulsioxímetro no funciona con inteligencia artificial. Pero aun así, no presta suficiente atención a una población concreta. Cuando se emplea la IA, debemos intentar comprender sus errores, no para perdonarlos, sino para corregirlos. Los prejuicios afectan a todos los aspectos de la sociedad humana y, en todos ellos, merecen una respuesta seria.

Otra fuente de identificación errónea es la rigidez. Consideremos el caso de un animal que se identifica erróneamente como una pistola. La imagen induce a error a las IA porque contiene características sutiles que los humanos no detectan, pero las IA sí, y pueden confundirlas. La IA no posee lo que llamamos sentido común. A veces confunde dos objetos que los humanos podrían

distinguir rápida y fácilmente. A menudo, lo que confunde (y cómo lo confunde) es inesperado, entre otras cosas porque, en el momento de escribir estas líneas, la solidez de los regímenes de auditoría y cumplimiento de la IA es escasa. En el mundo real, un fallo inesperado puede ser más perjudicial, o al menos más difícil, que uno esperado: la sociedad no puede mitigar lo que no prevé.

La fragilidad de la IA es un reflejo de la superficialidad de lo que aprende. Las asociaciones entre aspectos de entradas y salidas basadas en el aprendizaje supervisado o de refuerzo son muy diferentes de la verdadera comprensión humana, con sus muchos grados de conceptualización y experiencia. La fragilidad es también un reflejo de la falta de autoconciencia de las IA. Una IA no es sensible. No sabe lo que no sabe. Por consiguiente, no puede identificar y evitar lo que para los humanos podrían ser errores obvios. Esta incapacidad de la IA para comprobar por sí misma errores que, de otro modo, serían evidentes, subraya la importancia de desarrollar pruebas que permitan a los humanos identificar los límites de las capacidades de una IA, revisar sus líneas de acción propuestas y predecir cuándo es probable que una IA falle.

En consecuencia, es vital desarrollar procedimientos para evaluar si una IA funcionará como se espera. Como el aprendizaje automático será el motor de la IA en un futuro previsible, los humanos seguiremos sin saber qué está aprendiendo una IA y cómo sabe lo que ha aprendido. Aunque esto pueda resultar desconcertante, no debería serlo: el aprendizaje humano suele ser igualmente opaco. Artistas y deportistas, escritores y mecánicos, padres e hijos —de hecho, todos los seres humanos— actúan a menudo por intuición y, por tanto, son incapaces de articular qué o cómo han aprendido. Para hacer frente a esta opacidad, las sociedades han desarrollado innumerables programas de certificación profesional, reglamentos y leyes. Técnicas similares deberían aplicarse a las IA; por ejemplo, las sociedades podrían permitir el empleo de una IA solo después de que sus creadores demuestren su fiabilidad mediante procesos de prueba. El desarrollo

de programas de certificación profesional, control del cumplimiento y supervisión de la IA —y de los conocimientos de auditoría que requerirá su ejecución— será un proyecto social crucial.

En la industria, las pruebas previas al uso existen en un espectro. Los desarrolladores de aplicaciones a menudo se apresuran a comercializar los programas, corrigiendo los fallos en tiempo real, mientras que las empresas aeroespaciales hacen lo contrario: prueban religiosamente sus aviones antes de que un solo cliente ponga un pie a bordo. La variación de estos regímenes depende de varios factores, sobre todo del riesgo inherente a la actividad. A medida que se multipliquen las implementaciones de IA, los mismos factores —riesgo inherente, supervisión reglamentaria, fuerzas del mercado— probablemente los distribuirán en el mismo espectro, con las IA que conducen automóviles sometidas a una supervisión significativamente mayor que las IA que impulsan plataformas de redes para el entretenimiento y la conexión, como TikTok.

La división entre las fases de aprendizaje e inferencia en el aprendizaje automático permite que funcione un régimen de pruebas como este. Cuando una IA aprende continuamente, incluso mientras funciona, puede desarrollar comportamientos inesperados o indeseables, como le ocurrió a Tay, el *chatbot* de Microsoft, en 2016. En Internet, Tay se encontró con un discurso de odio y rápidamente comenzó a imitarlo, lo que obligó a sus creadores a cerrarlo. La mayoría de las IA, sin embargo, se entrenan en una fase distinta de la fase operativa: sus modelos aprendidos —los parámetros de sus redes neuronales— son estáticos cuando salen del entrenamiento. Como la evolución de una IA se detiene tras el entrenamiento, los humanos pueden evaluar sus capacidades sin temor a que desarrolle comportamientos inesperados e indeseados una vez completadas las pruebas. En otras palabras, cuando el algoritmo es fijo, un coche autónomo entrenado para detenerse en los semáforos en rojo no puede «decidir» de repente empezar a saltárselos. Esta propiedad hace posible la realización de pruebas y certificaciones exhaustivas: los ingenieros pueden examinar el comportamiento de una IA

autodirigida en un entorno seguro antes de cargarla en un coche, donde un error podría costar vidas. Por supuesto, la fijeza del algoritmo no significa que una IA no vaya a comportarse de forma inesperada cuando se ponga en contextos nuevos, simplemente significa que es posible realizar pruebas previas. La auditoría de los conjuntos de datos es otro control de calidad: asegurándose de que una IA de reconocimiento facial se entrena en diversos conjuntos de datos, o de que un *chatbot* se entrena en conjuntos de datos libres de discursos de odio, los desarrolladores pueden reducir aún más el riesgo de que la IA falle cuando sea operativa.

En el momento de escribir estas líneas, la IA está limitada por su código de tres maneras. En primer lugar, el código establece los parámetros de las posibles acciones de la IA. Estos parámetros pueden ser muy amplios, lo cual permite un amplio margen de autonomía y, por tanto, de riesgo. Una IA autodirigida puede frenar, acelerar y girar, y cualesquiera de estas acciones podría precipitar una colisión. No obstante, los parámetros del código establecen algunos límites al comportamiento de la IA. Aunque AlphaZero desarrolló estrategias de ajedrez novedosas, no lo hizo rompiendo las reglas del ajedrez; no movió peones hacia atrás de repente. Las acciones fuera de los parámetros del código están fuera del vocabulario de la IA. Y si el programador no pone ahí la capacidad, o prohíbe explícitamente la acción, la IA no puede hacerla. En segundo lugar, la IA está limitada por su función objetivo, que define y asigna lo que debe optimizar. En el caso del modelo que descubrió la halicina, la función objetivo era la relación entre las propiedades químicas de las moléculas y su potencial antibiótico. Limitada por su función objetivo, esa IA no podría haber tratado de identificar moléculas que pudieran, por ejemplo, ayudar a curar el cáncer. Por último, y lo que es más obvio, la IA solo puede procesar los datos que está diseñada para reconocer y analizar. Sin la intervención humana en forma de programa auxiliar, una IA traductora no puede evaluar imágenes: los datos le parecerían absurdos.

Algún día, las IA podrán escribir su propio código. Por ahora, los esfuerzos por diseñar este tipo de IA son incipientes y especulativos. Pero incluso entonces, las IA no serían autorreflexivas; sus funciones objetivo seguirían definiéndolas. Podrían escribir código de la misma forma que AlphaZero juega al ajedrez: brillantemente, pero sin reflexión ni voluntad, con estricto apego a las reglas.

Hacia dónde va la IA

Los avances en los algoritmos de aprendizaje automático, combinados con el aumento de los datos y la potencia computacional, han permitido un rápido progreso en la aplicación de la IA, captando imaginaciones y dólares de inversión. La explosión de la investigación, el desarrollo y la comercialización de la IA, especialmente del aprendizaje automático, es mundial, pero se ha concentrado en gran medida en Estados Unidos y China.⁶ Universidades, laboratorios, *start-ups* y conglomerados de ambos países han estado a la vanguardia del desarrollo y la aplicación del aprendizaje automático a problemas cada vez más numerosos y complejos.

Dicho esto, todavía hay que desarrollar y comprender muchos aspectos de la IA y el aprendizaje automático. La IA basada en el aprendizaje automático requiere muchos datos de entrenamiento. Los datos de entrenamiento, a su vez, requieren una infraestructura informática considerable; por lo tanto, el reentrenamiento de la IA resulta prohibitivamente caro, aunque sea deseable hacerlo. Debido a que los requisitos de datos y computación limitan el desarrollo de una IA más avanzada, idear métodos de entrenamiento que utilicen menos datos y menos potencia informática es una frontera crítica.

Además, a pesar de los grandes avances en el aprendizaje automático, las actividades complejas que requieren sintetizar varias tareas siguen siendo un reto para la IA. Conducir un coche, por ejemplo, ha demostrado ser un reto formidable, pues requiere la realización de funciones que van desde la percepción visual a la

navegación, pasando por la evitación proactiva de accidentes, todo ello de manera simultánea. Aunque este campo ha avanzado enormemente en la última década, los escenarios de conducción varían significativamente en cuanto al reto que supone alcanzar un rendimiento de nivel humano. En la actualidad, las IA pueden obtener buenos resultados en entornos estructurados, como autopistas de acceso limitado y calles suburbanas con pocos peatones o ciclistas. Sin embargo, el funcionamiento en entornos caóticos como el tráfico en hora punta de una ciudad sigue siendo un reto. La conducción en autopista es especialmente interesante, porque los conductores humanos suelen aburrirse y distraerse, lo cual posibilita que en un futuro no muy lejano las IA sean más seguras que los conductores humanos en los viajes de larga distancia.

Predecir el ritmo de avance de la IA será difícil. En 1965, el ingeniero Gordon Moore predijo que la potencia de cálculo se duplicaría cada dos años, una previsión que ha demostrado ser notablemente duradera. Pero el progreso de la IA es mucho menos predecible. La IA de traducción de idiomas se estancó durante décadas y, después, gracias a una confluencia de técnicas y potencia de cálculo, avanzó a un ritmo vertiginoso. En pocos años, los humanos han desarrollado IA con una capacidad de traducción similar a la de un ser humano bilingüe. No se puede predecir con exactitud cuánto tardará la IA en alcanzar las cualidades de un traductor profesional dotado, si es que alguna vez lo hace.

Predecir con qué rapidez se aplicará la IA a otros campos es igualmente difícil. Pero podemos seguir esperando aumentos espectaculares de las capacidades de estos sistemas. Aunque estos avances tarden cinco, diez o veinticinco años, en algún momento se producirán. Las aplicaciones de IA existentes serán más compactas, eficaces, baratas y, por tanto, se utilizarán con más frecuencia. La IA formará parte cada vez más de nuestra vida cotidiana, tanto de forma visible como invisible.

Es razonable esperar que, con el tiempo, la IA progrese al menos tan rápido como lo ha hecho la potencia de cálculo, y se multiplique por un millón en quince o veinte años. Este progreso permitirá la creación de redes neuronales que, en escala, iguallen al cerebro humano. En el momento de escribir estas líneas, los transformadores generativos tienen las redes más grandes. El GPT-3 tiene unos 10¹² pesos de este tipo. Pero hace poco la Academia de las Ciencias de Pekín, financiada por el Estado, anunció un modelo generativo del lenguaje con 10 veces más pesos que el GPT-3. Esto sigue siendo 10⁸ veces menos que las estimaciones de las sinapsis del cerebro humano. Pero si los avances avanzan al ritmo de duplicarse cada dos años, esta brecha podría cerrarse en menos de una década. Por supuesto, la escala no se traduce directamente en inteligencia. De hecho, se desconoce el nivel de capacidad que mantendrá una red. Algunos primates tienen cerebros de tamaño similar o incluso superior al humano, pero no muestran nada que se acerque a la perspicacia humana. Es probable que el desarrollo produzca «sabios» de la IA, programas capaces de superar con creces el rendimiento humano en áreas específicas, como los campos científicos avanzados.

El sueño de la inteligencia artificial general

Algunos desarrolladores están ampliando las fronteras de las técnicas de aprendizaje automático para crear lo que se ha dado en llamar inteligencia artificial general (AGI), por sus siglas en inglés. Al igual que la IA, la AGI no tiene una definición precisa. Sin embargo, suele entenderse como una IA capaz de llevar a cabo cualquier tarea intelectual que el ser humano sea capaz de realizar, a diferencia de la IA «restringida» actual, desarrollada para realizar una tarea específica.

Incluso más que para la IA actual, el aprendizaje automático es fundamental para el desarrollo de la AGI, aunque las limitaciones prácticas pueden limitar el alcance de sus conocimientos a un

número discreto de campos, del mismo modo que el ser humano más completo debe especializarse. Una posible vía para el desarrollo de la inteligencia artificial consiste en entrenar a las IA tradicionales en varios campos y combinar sus conocimientos en una única IA. Una AGI de este tipo podría ser más completa, capaz de realizar un conjunto más amplio de actividades, y menos frágil, al cometer errores menos drásticos en los límites de sus conocimientos.

Sin embargo, científicos y filósofos discrepan sobre si la verdadera inteligencia artificial es posible y sobre qué características podría tener. Si la AGI es posible, ¿tendrá las capacidades de un humano medio o del mejor humano en un campo determinado? En cualquier caso, incluso si la AGI pudiera desarrollarse del modo descrito anteriormente —combinando IA tradicionales, entrenándolas estrecha y profundamente y aglomerándolas gradualmente para desarrollar una base de conocimientos más amplia—, plantearía retos incluso a los investigadores mejor financiados y más sofisticados. Desarrollar este tipo de IA requeriría una enorme capacidad de cálculo y sería extremadamente caro —con la tecnología actual, del orden de miles de millones—, por lo que pocos podrían permitirse crearlas.

En cualquier caso, no está claro que la creación de la inteligencia artificial vaya a alterar sustancialmente la trayectoria que los algoritmos de aprendizaje automático han marcado para la humanidad. Tanto si se trata de IA como de AGI, los desarrolladores humanos seguirán desempeñando un papel importante en la creación y el funcionamiento. Los algoritmos, los datos de entrenamiento y los objetivos del aprendizaje automático los determinan las personas que desarrollan y entrenan la IA, por lo que reflejan los valores, las motivaciones, los objetivos y el juicio de esas personas. Aunque las técnicas de aprendizaje automático se vuelvan más sofisticadas, estas limitaciones persistirán.

Independientemente de que la IA siga siendo limitada o se generalice, cada vez será más frecuente y potente. A medida que disminuyan los costes de desarrollo y despliegue, los dispositivos

automatizados que funcionen con IA serán cada vez más accesibles. De hecho, ya existen interfaces conversacionales como Alexa, Siri y Google Assistant. Los vehículos, las herramientas y los electrodomésticos estarán cada vez más equipados con IA que automatizará su actividad bajo nuestra dirección y supervisión. Las IA se integrarán en aplicaciones de dispositivos digitales e Internet, guiarán las experiencias de los consumidores y revolucionarán las empresas. El mundo que conocemos será a la vez más automático y más interactivo (entre humanos y máquinas), aunque no esté poblado de los robots polivalentes de las películas de ciencia ficción. En los resultados más sorprendentes, se salvarán vidas humanas. Los vehículos autónomos reducirán el número de muertes; otras IA identificarán enfermedades antes y con mayor precisión; otras descubrirán fármacos y métodos de administración de fármacos con costes de investigación más reducidos, lo cual dará lugar, esperamos, al desarrollo de tratamientos para males crónicos y curas para enfermedades raras. Los aviadores de IA pilotarán o copilotarán flotas de drones de reparto e incluso aviones de combate. Los programadores de IA completarán los programas esbozados por los desarrolladores humanos; los redactores de IA completarán los anuncios concebidos por los vendedores humanos. Potencialmente, la eficiencia del transporte y la logística aumentará de forma espectacular. La IA reducirá el consumo de energía y probablemente encontrará otras formas de moderar el impacto medioambiental humano. En los ámbitos de la paz y la guerra, sus efectos materiales serán sorprendentes.

Pero sus repercusiones sociales son difíciles de predecir. Pensemos en la traducción de idiomas. La traducción universal del lenguaje hablado y del texto facilitará la comunicación como nunca antes. Esta traducción impulsará el comercio y permitirá un intercambio intercultural sin precedentes. Sin embargo, esta nueva capacidad también conllevará nuevos retos. Del mismo modo que las redes sociales no solo permiten el intercambio de ideas, sino que fomentan también la polarización, promulgan la desinformación y

difunden discursos de odio, la traducción automática puede acercar lenguas y culturas con efectos explosivos. Durante siglos, los diplomáticos han gestionado cuidadosamente el contacto intercultural para evitar ofensas accidentales, del mismo modo que la sensibilización cultural ha acompañado a menudo a la formación lingüística. La traducción instantánea elimina estos amortiguadores. Las sociedades pueden encontrarse sin querer ofendiendo y siendo ofendidas. Al depender de la traducción automática, ¿se esforzará menos la gente por comprender otras culturas y naciones, aumentando su tendencia natural a ver el mundo a través de la lente de su propia cultura? ¿O se sentirán más intrigados por otras culturas? ¿Puede la traducción automática reflejar de algún modo las diferentes historias culturales y sensibilidades? Probablemente no haya una respuesta única.

Las IA más avanzadas requieren una gran cantidad de datos, una enorme potencia de cálculo y técnicos cualificados. No es de extrañar que las organizaciones con acceso a tales recursos, tanto comerciales como gubernamentales, impulsen gran parte de la innovación en este nuevo campo. Y a los líderes les llegan más recursos. Así, un ciclo de concentración y avance ha definido la IA, y ha dado forma a la experiencia de individuos, empresas y naciones. En muchos ámbitos, desde la comunicación al comercio, pasando por la seguridad y la propia conciencia humana, la IA transformará nuestras vidas y nuestro futuro. Todos debemos asegurarnos de que la IA no se crea de forma aislada y, en consecuencia, hemos de prestar atención tanto a sus posibles beneficios como a sus potenciales riesgos.

¹ Véase Eric Schmidt y Jared Cohen, *The New Digital Age: Reshaping the Future of People, Nations, and Business* (New York: Alfred A. Knopf, 2013).

² Alan Turing, «Computing Machinery and Intelligence», *Mind* 59, nº 236 (October, 1950), 433-460, reimpresso en B. Jack Copeland, ed., *The Essential Turing: Seminal Writings in Computing, Logic, Philosophy, Artificial Intelligence, and Artificial Life*

Plus the Secrets of Enigma (Oxford, Reino Unido: Oxford University Press, 2004), 441-464.

³ Específicamente, una búsqueda en el árbol de búsqueda Monte Carlo de movimientos futuros habilitados o excluidos.

⁴ James Vincent, «Google 'Fixed' Its Racist Algorithm by Removing Gorillas from Its Image-Labeling Tech», The Verge, January, 12, 2018, <https://www.theverge.com/2018/1/12/16882408/google-racist-gorillas-photo-recognition-algorithm-ai>.

⁵ James Vincent, «Google's AI Thinks This Turtle Looks Like a Gun, Which Is a Problem», The Verge, November 2, 2017, <https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed>.

⁶ Y, en menor medida, Europa y Canadá.

⁷ Sin embargo, algunos episodios históricos ofrecen paralelismos instructivos. Para un estudio de las interacciones entre el poder centralizado y las redes, véase Niall Ferguson, *La plaza y la torre*, editorial Debate, Barcelona, 2018.

⁸ Aunque el término plataforma puede utilizarse para referirse a muchas cosas diferentes en el ámbito digital, utilizamos plataforma de redes específicamente para referirnos a los servicios en línea con efectos de red positivos.

Plataformas de redes globales

Las visiones de ficción sobre el futuro de la tecnología de la IA suelen evocar imágenes de elegantes coches autónomos totalmente automatizados y robots sensibles que coexisten con los humanos en hogares y lugares de trabajo, y conversan con sus usuarios con una inteligencia asombrosa. Inspiradas en este tipo de escenas de ciencia ficción, las concepciones populares de la IA a menudo implican máquinas que desarrollan una aparente autoconciencia, lo que inevitablemente las lleva a malinterpretar, negarse a obedecer o, finalmente, levantarse contra sus creadores humanos. Pero los temores subyacentes tras estas fantasías comunes confunden la cuestión al suponer que la culminación de la IA será actuar como seres humanos individuales. Nos convendría más reconocer que la IA ya está a nuestro alrededor —a menudo de formas no del todo evidentes— y reorientar nuestras ansiedades tecnológicas hacia el fomento de una mayor comprensión y transparencia respecto a la integración de la IA en nuestras vidas.

Las redes sociales, las búsquedas en Internet, los vídeos en *streaming*, la navegación, los viajes compartidos y otros muchos servicios en línea no podrían funcionar como lo hacen sin el uso cada vez más extendido de la IA. Al utilizar estos servicios en línea para las actividades básicas de la vida diaria —para ofrecer recomendaciones de productos y servicios, seleccionar rutas, establecer conexiones sociales y obtener ideas o respuestas—, personas de todo el mundo participan en un proceso a la vez mundano y revolucionario. Confiamos en la IA para que nos ayude en nuestras tareas cotidianas sin entender necesariamente cómo o

por qué funciona en un momento dado. Estamos formando nuevos tipos de relaciones que tendrán implicaciones sustanciales para los individuos, las instituciones y las naciones: entre la IA y las personas, entre las personas que utilizan servicios facilitados por la IA y entre los creadores y operadores de estos servicios y los gobiernos.

Sin grandes alardes —ni siquiera visibilidad— estamos integrando la inteligencia no humana en el tejido básico de la actividad humana. Esto se está produciendo rápidamente, relacionado con un nuevo tipo de entidad que denominamos «plataformas de redes»: servicios digitales que proporcionan valor a sus usuarios mediante la agregación de esos usuarios en grandes cantidades, a menudo a escala transnacional y mundial. A diferencia de la mayoría de los productos y servicios, cuyo valor para cada usuario es independiente de la presencia de otros usuarios, o incluso se ve disminuido por ella, el valor y el atractivo de una plataforma de redes aumenta a medida que más usuarios la adoptan, un proceso que los economistas denominarían efecto de red positivo. A medida que más usuarios se sienten atraídos por las plataformas seleccionadas, estas agrupaciones tienden a dar lugar a un pequeño número de proveedores que ofrecen un servicio determinado, cada uno con una gran base de usuarios, a veces cientos de millones o incluso miles de millones. Estas plataformas de redes se basan cada vez más en la IA, y producen una intersección entre los seres humanos y la IA a una escala que sugiere un acontecimiento clave para nuestra civilización.

A medida que la IA asume mayores funciones en plataformas de redes más variadas, las manifestaciones básicas de estas plataformas se están convirtiendo en material para titulares y maniobras geopolíticas, mientras configuran aspectos de la realidad cotidiana de los individuos. Sin medios adicionales de explicación, debate y supervisión compatibles con los valores de una sociedad y capaces de conducir a cierto grado de consenso social y político, puede desencadenarse una rebelión contra la llegada de fuerzas

nuevas y aparentemente impersonales e inexorables, como ocurrió con el auge del Romanticismo en el siglo XIX y la explosión de ideologías radicales en el XX. Antes de que se produzca una perturbación significativa, los gobiernos, los operadores de plataformas de redes y los usuarios deben considerar la naturaleza de sus objetivos, las premisas y parámetros básicos de sus interacciones y el tipo de mundo que pretenden crear.

En menos de una generación, las plataformas de redes de mayor éxito han reunido bases de usuarios mayores que las poblaciones de la mayoría de las naciones e incluso de algunos continentes. Sin embargo, las grandes poblaciones de usuarios reunidas en plataformas de redes populares tienen fronteras más difusas que las de la geografía política, y las plataformas de redes son operadas por partes con intereses que pueden diferir de los de una nación. Los operadores de las plataformas de redes no piensan necesariamente en términos de prioridades gubernamentales o de estrategia nacional, sobre todo si esas prioridades y estrategias pueden entrar en conflicto con el servicio a sus clientes. Estas plataformas de redes pueden albergar o facilitar interacciones económicas y sociales que superan (en número y escala) a las de la mayoría de los países, a pesar de que las plataformas no hayan formulado una política económica o social como la que tendría un gobierno. Así, aunque funcionan como entidades comerciales, algunas plataformas de redes se están convirtiendo en actores geopolíticamente significativos, en virtud de su escala, función e influencia.

Muchas de las plataformas de redes más importantes tienen su origen en Estados Unidos (Google, Facebook, Uber) o China (Baidu, WeChat, Didi Chuxing). Como resultado, estas plataformas de redes tratan de construir sus bases de usuarios y asociaciones comerciales en regiones que contienen mercados comercial y estratégicamente importantes para Washington y Pekín. Esta dinámica introduce nuevos factores en los cálculos de política exterior. La competencia comercial entre plataformas de redes puede afectar a la competencia geopolítica entre gobiernos, y llegar a veces a

encabezar la agenda de la diplomacia. Esto se complica aún más por el hecho de que las culturas y estrategias corporativas de los operadores de plataformas de redes se desarrollan a menudo para reflejar las prioridades de los clientes y de los centros de investigación y tecnología, que pueden estar lejos de las capitales nacionales.

En los países en los que operan, ciertas plataformas de redes se han convertido en parte integrante de la vida individual, del discurso político nacional, del comercio, de la organización empresarial e incluso de las funciones gubernamentales. Sus servicios —aun los que no existían en forma alguna hasta hace poco— parecen ahora indispensables. Como entidad sin un solo precedente directo de épocas anteriores, las plataformas de redes se sitúan a veces en una relación ambigua con normas y expectativas desarrolladas en gran medida en un mundo predigital.

La cuestión de cómo las plataformas de redes establecen las normas comunitarias —las reglas, fijadas por cada operador (a menudo administradas con la ayuda de la IA), que rigen qué contenidos está permitido crear y compartir— ofrece un ejemplo claro de la incongruencia entre el espacio digital moderno y las normas y expectativas tradicionales. Aunque en principio la mayoría de las plataformas de redes son ajenas a los contenidos, en algunas situaciones sus normas comunitarias llegan a ser tan influyentes como las leyes nacionales. El contenido que una plataforma de redes y su IA permiten o favorecen puede ganar rápidamente prominencia; el contenido que disminuyen o a veces incluso prohíben puede quedar relegado a la oscuridad. El material que se determina que contiene desinformación o viola otras normas de contenido puede ser retirado de la circulación pública.

Cuestiones como estas han surgido con rapidez, en parte porque las plataformas de redes (y su IA) se han expandido rápidamente en un mundo digital que trasciende la geografía. Estas plataformas conectan a grandes grupos de usuarios a través del espacio y el tiempo —con agregaciones de datos accesibles instantáneamente—

de una forma a la que pocas creaciones humanas se aproximan.¹ Para agravar el reto, una vez que la IA ha sido entrenada, suele actuar más rápido que la velocidad de la cognición humana. Estos fenómenos no son intrínsecamente ni positivos ni negativos; son realidades ocasionadas por los problemas que los seres humanos tratamos de resolver, las necesidades que deseamos satisfacer y la tecnología que creamos para servir a nuestros fines. Estamos experimentando y facilitando cambios que requieren nuestra atención —en el pensamiento, la cultura, la política y el comercio— mucho más allá del alcance de una sola mente humana o de un producto o servicio concreto.

Cuando el mundo digital comenzó a expandirse hace décadas, no se esperaba que los creadores desarrollaran o debieran desarrollar un marco filosófico o definir su relación fundamental con los intereses nacionales o mundiales. Al fin y al cabo, por lo general no se habían hecho tales afirmaciones en otras industrias. En su lugar, la sociedad y los gobiernos evaluaban los productos y servicios digitales a partir de lo que funcionaba. Los ingenieros buscaban soluciones prácticas y eficientes: conectar a los usuarios con la información y los espacios sociales en línea, a los pasajeros con los coches y los conductores, y a los clientes con los productos. Había entusiasmo general por las nuevas capacidades y oportunidades. Había poca demanda de predicciones sobre cómo estas soluciones virtuales podrían afectar a los valores y el comportamiento de sociedades enteras, como los patrones de uso de vehículos y la congestión del tráfico con los viajes compartidos o los alineamientos políticos y geopolíticos del mundo real de las instituciones nacionales con los medios sociales.

Las plataformas de redes dotadas de IA se crearon incluso más recientemente; con menos de una década de desarrollo, aún no hemos establecido ni siquiera el vocabulario y los conceptos básicos para un debate informado sobre esta tecnología, una laguna que este libro pretende contribuir a remediar. Diversos individuos, empresas, partidos políticos, organizaciones cívicas y gobiernos

tendrán inevitablemente opiniones divergentes sobre el funcionamiento y la regulación adecuados de las plataformas de redes dotadas de IA. Lo que parece intuitivo para el ingeniero de software puede resultar desconcertante para el líder político o inexplicable para el filósofo. Lo que el consumidor ve como una comodidad, el responsable de seguridad nacional puede considerarlo una amenaza inaceptable o el líder político puede rechazarlo por no ajustarse a los objetivos nacionales. Lo que una sociedad considera una garantía, otra lo interpreta como una pérdida de elección o de libertad.

La naturaleza y la escala de las plataformas de redes están uniendo las perspectivas y prioridades de mundos diferentes en alineaciones complejas, creando a veces tensiones y perplejidad mutua. Para que los agentes individuales, nacionales e internacionales lleguen a conclusiones fundamentadas sobre su relación con la IA —y entre sí—, debemos buscar un marco de referencia común, empezando por establecer los términos de unos debates políticos fundamentados. Incluso si nuestras concepciones difieren, debemos tratar de entender las plataformas de redes habilitadas por IA y evaluar sus implicaciones para los individuos, las empresas, las sociedades, las naciones, los gobiernos y las regiones. Debemos actuar urgentemente en cada nivel.

Comprender las plataformas de redes

Las plataformas de redes son fenómenos intrínsecamente a gran escala. Una de las características definitorias de una plataforma de redes es que, a medida que aumenta el número de personas a quienes presta servicio, se hace más útil y deseable para los usuarios.² La IA es cada vez más importante para las plataformas de redes que pretenden prestar sus servicios a gran escala y, como resultado, casi todos los usuarios de Internet se encuentran hoy con la IA, o al menos con contenidos en línea creados por la IA, numerosas veces al día.

Por ejemplo, Facebook (como muchas otras redes sociales) ha desarrollado normas comunitarias cada vez más específicas para la eliminación de contenidos y cuentas censurables, enumerando docenas de categorías de contenidos prohibidos a finales de 2020. Como la plataforma cuenta con miles de millones de usuarios activos mensuales y miles de millones de visitas diarias,³ la magnitud de la supervisión de contenidos en Facebook supera las capacidades de los moderadores humanos por sí solos. A pesar de que, al parecer, Facebook tiene decenas de miles de personas trabajando en la moderación de contenidos —con el objetivo de eliminar el contenido ofensivo antes de que los usuarios lo vean—, la escala es simplemente tal que no se puede lograr sin IA. Estas necesidades de supervisión en Facebook y otras empresas han impulsado una amplia labor de investigación y desarrollo en un esfuerzo por automatizar el análisis de textos e imágenes mediante la creación de técnicas cada vez más sofisticadas de aprendizaje automático, procesamiento del lenguaje natural y visión por ordenador.

En el caso de Facebook, el número de eliminaciones es actualmente del orden de unos mil millones de cuentas falsas y mensajes de spam al trimestre, así como decenas de millones de contenidos que incluyen desnudos o actividad sexual, acoso, explotación, incitación al odio, drogas y violencia. Para llevar a cabo estas eliminaciones con precisión, a menudo es necesario el juicio humano. Así pues, en su mayor parte, los operadores humanos y los usuarios de Facebook confían en la IA para determinar qué contenidos merecen ser consumidos o revisados.⁴ Aunque solo una pequeña parte de las retiradas son recurridas, las que lo son tienden a ser retiradas automatizadas.

Del mismo modo, la IA desempeña un papel importante en el motor de búsqueda de Google, pero un papel relativamente reciente y en rápida evolución. Al principio, el motor de búsqueda de Google se basaba en algoritmos muy complejos desarrollados por humanos para organizar, clasificar y guiar a los usuarios hacia la información. Estos algoritmos consistían en un conjunto de reglas para gestionar

las posibles consultas de los usuarios. Cuando los resultados no resultaban útiles, los desarrolladores humanos podían ajustarlos. En 2015, el equipo de búsqueda de Google pasó de utilizar estos algoritmos desarrollados por humanos a aplicar el aprendizaje automático. Este cambio marcó un antes y un después: la incorporación de la IA ha mejorado enormemente la calidad y la usabilidad del motor de búsqueda, y lo ha hecho más capaz de anticiparse a las preguntas y organizar resultados precisos. Sin embargo, a pesar de las mejoras significativas en el motor de búsqueda de Google, los desarrolladores tenían una comprensión relativamente vaga de por qué las búsquedas producían resultados particulares. Los humanos aún pueden guiar y ajustar el motor de búsqueda, pero no pueden explicar por qué una página en particular está mejor clasificada que otra. Para lograr una mayor comodidad y precisión, los desarrolladores humanos han tenido que renunciar voluntariamente a cierto grado de comprensión directa.⁵

Como ilustran estos ejemplos, las principales plataformas de redes dependen cada vez más de la IA para prestar servicios, satisfacer las expectativas de los clientes y cumplir diversos requisitos gubernamentales. A medida que la IA se vuelve cada vez más crítica para el funcionamiento de las plataformas de redes, también se está convirtiendo, de forma gradual y discreta, en un clasificador y moldeador de la realidad y, de hecho, en un actor en el escenario nacional y mundial.

El potencial de influencia social, económica, política y geopolítica de cada gran plataforma de redes (y su IA) aumenta sustancialmente por su grado de efectos de red positivos. Los efectos de red positivos se producen en las actividades de intercambio de información en las que el valor aumenta con el número de participantes. Cuando el valor aumenta de esta manera, el éxito tiende a producir más éxito y una mayor probabilidad de predominio final. La gente gravita de forma natural hacia las concurrencias existentes, lo cual conduce a mayores agregaciones de usuarios. Para una plataforma de redes relativamente libre de

fronteras, esta dinámica conduce a un ámbito geográfico más amplio, a menudo transnacional, con la correspondiente reducción de servicios competidores importantes.

Los efectos de red positivos no se originaron con las plataformas de redes. Sin embargo, antes del auge de la tecnología digital, la aparición de tales efectos era relativamente rara. De hecho, un aumento del número de usuarios puede fácilmente restarle valor a un producto o servicio tradicional en lugar de añadirsele. Esta situación puede producir escasez (en el caso de un producto o servicio muy demandado o agotado), retrasos (en el caso de un producto o servicio que no puede entregarse simultáneamente a todos los clientes que lo desean) o una pérdida del sentido de exclusividad que dio a un producto su caché inicial (por ejemplo, un artículo de lujo que se vuelve menos codiciado cuando está ampliamente disponible).

El ejemplo clásico de efectos de red positivos surge en los propios mercados, ya sean de bienes o de valores. Al menos desde principios del siglo XVII, los comerciantes de acciones y bonos de las Indias Orientales Holandesas se reunían en Ámsterdam, donde las bolsas de valores proporcionaban un medio para que compradores y vendedores llegaran a una valoración común con el fin de negociar valores. Con la participación activa de más compradores y vendedores, una bolsa de valores se vuelve más útil y valiosa para los participantes individuales. Contar con más participantes aumenta las posibilidades de que se produzca una transacción y de que su valoración sea «exacta», porque la transacción refleja un mayor número de negociaciones individuales entre compradores y vendedores. Una vez que una bolsa ha reunido una masa crítica de usuarios en un mercado determinado, tiende a convertirse en la primera parada para nuevos compradores y vendedores, y deja pocos incentivos u oportunidades para que otra bolsa compita en la oferta de exactamente el mismo servicio.

Cuando se desarrollaron los teléfonos tradicionales, las redes telefónicas también mostraron fuertes efectos de red positivos. Para

un servicio telefónico que depende de cables físicos para conectar las llamadas, tener un mayor número de abonados en la misma red crea un mayor valor para cada abonado. Así, en los primeros tiempos de la telefonía, hubo un fuerte crecimiento para los grandes proveedores de servicios. En Estados Unidos, la universalidad se logró inicialmente con una red muy grande operada por AT&T (originalmente Bell Telephone), interconectada a una serie de proveedores más pequeños, en su mayoría rurales. En la década de 1980, los avances tecnológicos permitieron a los proveedores de servicios telefónicos conectarse más fácilmente entre sí, posibilitando así a los abonados de los nuevos proveedores llegar sin problemas a los de cualquier servicio (nacional). Estos avances facilitaron la disolución reglamentaria de AT&T, pues demostraron a los clientes que el valor seguiría siendo alto incluso sin un único gran proveedor. A medida que la tecnología seguía evolucionando, los clientes podían comunicarse con cualquiera desde sus teléfonos independientemente de sus proveedores, lo cual redujo enormemente el efecto de red positivo.⁶

No hay ninguna razón intrínseca para que la dinámica de los efectos positivos de la red se detenga en las fronteras nacionales o regionales, y las plataformas de redes suelen expandirse más allá de esas fronteras terrestres. Las distancias físicas y las diferencias nacionales o lingüísticas rara vez son obstáculos a la expansión: el mundo digital es accesible desde cualquier lugar con conexión a Internet, y los servicios de las plataformas de redes pueden prestarse normalmente en varios idiomas. Las principales limitaciones a la expansión son las impuestas por los gobiernos o quizá la incompatibilidad tecnológica (las primeras favorecen a veces a las segundas). Así, para cada tipo de servicio, como las redes sociales y el *streaming* de vídeo, suele haber un pequeño número de plataformas de redes globales, quizá complementadas por otras locales. Sus usuarios se benefician y contribuyen a un fenómeno nuevo y aún poco conocido: el funcionamiento de la inteligencia no humana a escala mundial.

Comunidad, vida cotidiana y plataformas de redes

El mundo digital ha transformado nuestra experiencia de la vida cotidiana. A lo largo del día, una persona se beneficia de enormes cantidades de datos y contribuye a ellos. El alcance de estos datos y las opciones para consumirlos son demasiado inmensos y variados para que las mentes humanas puedan procesarlos por sí solas. El individuo llega a confiar, a menudo instintiva o inconscientemente, en los procesos de software para organizar y seleccionar la información necesaria o útil, al elegir las noticias que quiere atender, las películas que quiere ver y la música que quiere escuchar basándose en una combinación de elecciones individuales previas y selecciones ampliamente populares. Experimentar esta selección automatizada puede ser tan sencillo y satisfactorio que solo se nota en su ausencia: por ejemplo, pruebe a leer noticias en el *feed* de Facebook de otra persona o a ver películas utilizando la cuenta de Netflix de otra persona.

Las plataformas de redes con IA han acelerado este proceso de integración y profundizado las conexiones entre las personas y nuestra tecnología digital. Con IA diseñada y entrenada para intuir y responder a preguntas y objetivos humanos, una plataforma de redes puede convertirse en guía, intérprete y registro de opciones que la mente humana antes gestionaba por sí misma (aunque con menos eficacia). Las plataformas de redes llevan a cabo estas tareas agregando información y experiencias a una escala mucho mayor de la que puede abarcar una sola mente humana o una sola vida, lo cual les permite producir respuestas y recomendaciones que pueden parecer increíblemente acertadas. Por ejemplo, a la hora de comprar unas botas de invierno, ni siquiera el comprador individual más aplicado evaluaría cientos de miles de compras nacionales y regionales de artículos similares, tendría en cuenta las tendencias meteorológicas recientes, el factor de la época del año, revisaría sus búsquedas anteriores comparables e investigaría los patrones de

envío antes de decidir qué par de botas sería la mejor compra. Una IA, sin embargo, podría muy bien evaluar todos estos factores.

Como resultado, las personas se relacionan a menudo con las plataformas de redes impulsadas por la IA de un modo en que no lo han hecho históricamente con otros productos, servicios o máquinas. A medida que el individuo interactúa con la IA y esta se adapta a sus preferencias (navegación por Internet y consultas de búsqueda, historial de viajes, nivel de ingresos aparente, conexiones sociales), comienza a formarse una especie de asociación tácita. El individuo pasa a confiar en estas plataformas para llevar a cabo una combinación de funciones tradicionalmente asignadas a empresas, gobiernos y otros seres humanos, hasta que la plataforma de redes se convierte en una combinación de servicio postal, grandes almacenes, conserje, confesor y amigo.

La relación entre un individuo, una plataforma de redes y sus otros usuarios es una novedosa combinación de vínculo íntimo y conexión remota. Las plataformas de redes habilitadas por IA ya revisan cantidades sustanciales de datos de los usuarios, muchos de los cuales son personales (como la ubicación, la información de contacto, las redes de amigos y socios y la información financiera y sanitaria). Los usuarios recurren a la IA como guía o facilitador de una experiencia personalizada. La precisión y agudeza de la IA se derivan de su capacidad para revisar y reaccionar a una agregación de cientos de millones de relaciones similares y billones de interacciones similares a través del espacio (la amplitud geográfica de la base de usuarios) y el tiempo (la agregación de usos anteriores). Los usuarios de la plataforma de redes y su IA entran en una especie de pacto, interactuando y aprendiendo unos de otros.

Al mismo tiempo, la IA de una plataforma de redes sigue una lógica no humana y, en muchos sentidos, inescrutable para los humanos. Por ejemplo, en la práctica, cuando una plataforma de redes con IA evalúa una imagen, una publicación en las redes sociales o una consulta de búsqueda, es posible que los humanos no entiendan exactamente cómo funciona la IA en esa situación

concreta. Aunque los ingenieros de Google saben que su función de búsqueda con IA produce resultados más claros que los que habrían logrado sin ella, no siempre pueden explicar por qué un resultado concreto se clasifica mejor que otro. En gran medida, la IA se juzga por la utilidad de sus resultados, no por el proceso utilizado para llegar a ellos. Esto supone un cambio de prioridades con respecto a épocas anteriores, cuando cada paso de un proceso mental o mecánico era experimentado por un ser humano (un pensamiento, una conversación, un proceso administrativo) o podía ser detenido, inspeccionado y repetido por seres humanos.

Por ejemplo, en gran parte del mundo industrializado ya se está olvidando la época en la que para viajar era necesario «obtener indicaciones», un proceso manual que podía implicar una llamada telefónica previa a la persona a la que se iba a visitar, la consulta de un mapa impreso de la ciudad o del estado y, en no pocas ocasiones, una parada en una gasolinera o tienda de conveniencia para validar una suposición o corregir un error. Ahora, el proceso del viaje se desarrolla de forma mucho más eficiente gracias al uso de aplicaciones de mapas para teléfonos inteligentes. Estas aplicaciones no solo pueden evaluar varias rutas posibles y el tiempo que llevaría cada una de ellas basándose en lo que «saben» sobre las pautas históricas del tráfico a esa hora del día, sino que también pueden tener en cuenta los accidentes y otros retrasos propios de ese día (incluidos los que se producen durante el trayecto), así como otros indicios (como las búsquedas de otros usuarios) de que el tráfico puede empeorar a lo largo de una ruta determinada durante el tiempo que el usuario tardará en recorrerla.

El paso de los atlas a los servicios de navegación en línea ha resultado tan cómodo que pocos se han parado a pensar qué cambio revolucionario se ha producido o cuáles pueden ser sus consecuencias. El individuo y la sociedad han ganado en comodidad al entablar una nueva relación con una plataforma de redes y su operador, accediendo y pasando a formar parte de un conjunto de datos en evolución (que incluye el seguimiento de la ubicación del

individuo, al menos mientras se utiliza la aplicación) y confiando en que la plataforma de redes y sus algoritmos produzcan resultados precisos. En cierto sentido, la persona que utiliza un servicio de este tipo no conduce sola, sino que forma parte de un sistema en el cual la inteligencia humana y la de las máquinas colaboran para guiar a un grupo de personas a través de sus rutas individuales.

Es probable que la prevalencia de este tipo de compañía constante de la IA vaya en aumento. A medida que sectores como la sanidad, la logística, el comercio minorista, las finanzas, las comunicaciones, los medios de comunicación, el transporte y el entretenimiento producen avances comparables —a menudo facilitados por plataformas de redes—, nuestra experiencia de la realidad cotidiana se está transformando.

Cuando los usuarios recurren a plataformas de redes dotadas de IA para que les ayuden a realizar tareas, se benefician de un tipo de recopilación y destilación de información que ninguna generación anterior ha experimentado. La escala, la potencia y la capacidad de estas plataformas para seguir patrones novedosos proporcionan a los usuarios individuales comodidades y capacidades sin precedentes. Al mismo tiempo, estos usuarios están entablando una forma de diálogo hombre-máquina que nunca antes había existido. Las plataformas de redes habilitadas por IA tienen la capacidad de moldear la actividad humana de formas que el usuario humano puede no entender claramente, o incluso definir o expresar con claridad. Esto plantea cuestiones esenciales: ¿Con qué función objetivo opera esa IA? ¿Quién la diseña y con qué parámetros normativos?

Las respuestas a estas y otras preguntas similares seguirán configurando las vidas y las sociedades en el futuro: ¿Quién opera y define los límites de estos procesos? ¿Qué impacto pueden tener en las normas e instituciones sociales? ¿Y quién, si alguien lo hace, tiene acceso a lo que percibe la IA? Si ningún ser humano puede comprender o ver los datos de forma individualizada ni acceder a todos los pasos del proceso —es decir, si el papel del ser humano se

limita a diseñar, supervisar y establecer parámetros generales para la IA—, ¿deberían estas limitaciones ser reconfortantes, inquietantes, o ambas cosas?

Empresas y naciones

Los diseñadores no iniciaron este camino con el claro objetivo de inventar plataformas de redes dotadas de IA, sino que estas surgieron de forma incidental, en función de los problemas que las empresas, los ingenieros y sus clientes trataban de resolver. Los operadores de plataformas de redes desarrollaron su tecnología para satisfacer determinadas necesidades humanas: conectaron a compradores y vendedores, a solicitantes y proveedores de información, y a grupos de individuos que compartían intereses u objetivos comunes. Desplegaron la IA para mejorar —o, cada vez más, habilitar— sus servicios y aumentar su capacidad para satisfacer las expectativas de los usuarios (y a veces de los gobiernos).

A medida que las plataformas de redes han ido creciendo y evolucionando, algunas han llegado a afectar a actividades y sectores de la sociedad mucho más allá de su objetivo original. Y, como ya se ha señalado, las personas han llegado a confiar a ciertas plataformas de redes impulsadas por IA información que dudarían en mostrar a un amigo o al gobierno, como registros exhaustivos de dónde han ido, qué han hecho (y con quién), y qué han buscado y visto.

La dinámica que permite el acceso a esos datos personales coloca a las plataformas de redes, a sus operadores y a la IA por ellos empleada en nuevas posiciones de influencia social y política. Especialmente durante una era de distanciamiento social y trabajo a distancia influenciada por la pandemia, las sociedades han llegado a confiar en algunas plataformas de redes habilitadas por IA como una especie de recurso esencial y pegamento social: un facilitador de la expresión, el comercio, la entrega de alimentos y el transporte. Estos cambios se han producido a una escala y a una velocidad tal que,

hasta ahora, han superado a la posibilidad de una comprensión y un consenso más amplios sobre el papel de estas plataformas de redes en la sociedad y en la escena internacional.

Como ha demostrado el reciente papel de las redes sociales en la transmisión y moderación de la información y la desinformación políticas, algunas plataformas de redes han asumido funciones tan significativas como para influir potencialmente en la conducción de la gobernanza nacional. Esta influencia ha surgido en efecto por accidente, sin que necesariamente se haya buscado o preparado adecuadamente. Sin embargo, las habilidades, instintos y percepciones conceptuales que producen la excelencia en el mundo de la tecnología inevitablemente no coinciden con las del mundo del gobierno. Cada esfera tiene su propio lenguaje, estructuras organizativas, principios animadores y valores fundamentales. Una plataforma de redes que opere de acuerdo con sus objetivos comerciales estándar y las demandas de sus usuarios puede, en efecto, estar trascendiendo al ámbito de la gobernanza y la estrategia nacional. A su vez, los gobiernos tradicionales pueden tener dificultades para discernir los motivos y las tácticas de la plataforma, incluso cuando tratan de ajustarlos a los objetivos nacionales y mundiales.

El hecho de que la IA funcione según sus propios procesos, diferentes y a menudo más rápidos que los procesos mentales humanos, añade otra complejidad. La IA desarrolla sus propios enfoques para cumplir las funciones objetivas especificadas. Produce resultados y respuestas que no son característicamente humanos y sí en gran medida independientes de las culturas nacionales o corporativas. La naturaleza global del mundo digital y la capacidad de la IA para supervisar, bloquear, adaptar, producir y distribuir información en plataformas de redes de todo el mundo importan estas complejidades al «espacio de la información» de sociedades dispares.

A medida que se utiliza una IA cada vez más sofisticada para habilitar las plataformas de redes, se configuran acuerdos sociales y

comerciales a escala nacional y mundial. Aunque las plataformas de redes sociales (y sus IA) suelen presentarse a sí mismas como agnósticas en cuanto a contenidos, no solo sus normas comunitarias, sino también su filtrado y presentación de la información pueden influir en la forma en que esta se crea, agrega y percibe. Cuando la IA recomienda contenidos y conexiones, clasifica información y conceptos y predice las preferencias y objetivos de los usuarios, puede reforzar inadvertidamente determinadas opciones individuales, grupales o sociales. En efecto, puede fomentar la distribución de ciertos tipos de información y la formación de ciertos tipos de conexiones, mientras desalienta otras. Esta dinámica puede afectar a los resultados sociales y políticos, independientemente de las intenciones de los operadores de la plataforma. Cada día, los usuarios individuales y los grupos se influyen mutuamente con rapidez y a gran escala a través de innumerables interacciones, sobre todo cuando están determinadas por complejas recomendaciones basadas en la inteligencia artificial; como resultado, los operadores pueden no tener una comprensión clara de lo que está ocurriendo en tiempo real. Y las complejidades se magnifican si el operador inyecta (consciente o inconscientemente) sus propios valores o propósitos.

Reconociendo estos retos, los intentos gubernamentales de abordar esta dinámica tendrán que proceder con sumo cuidado. Cualquier enfoque gubernamental de este proceso —ya sea para restringirlo, controlarlo o permitirlo— refleja necesariamente opciones y juicios de valor. Si un gobierno anima a las plataformas a etiquetar o bloquear determinados contenidos, o si exige a la IA que identifique y rebaje la información sesgada o «falsa», tales decisiones pueden operar efectivamente como motores de la política social con una amplitud e influencia únicas. En todo el mundo, la forma de abordar estas opciones se ha convertido en objeto de profundos debates, especialmente en las sociedades libres tecnológicamente avanzadas. Cualquiera sea el enfoque que se adopte, está garantizado que se producirá a una escala mucho

mayor que cualquier otra decisión jurídica o política del pasado, y con efectos potencialmente instantáneos en la vida cotidiana de millones o miles de millones de usuarios en muchas jurisdicciones gubernamentales.

La intersección entre la plataforma de redes y los ámbitos gubernamentales producirá resultados impredecibles y, en algunos casos, muy controvertidos. Sin embargo, en lugar de resultados claros, es más probable que lleguemos a una serie de dilemas con respuestas imperfectas. ¿Funcionarán los intentos de regular las plataformas de redes y sus IA en consonancia con los objetivos políticos y sociales de las distintas naciones (por ejemplo, reducir la delincuencia, combatir los prejuicios) y, en última instancia, producirán sociedades más justas? ¿O conducirán a gobiernos más poderosos e intrusivos que moldean los resultados a través de una máquina sustituta, cuya lógica es inefable y cuyas conclusiones resultan inevitables? En los intercambios iterativos que tienen lugar a lo largo del tiempo, entre continentes y entre bases de usuarios supranacionales, ¿las plataformas de redes habilitadas por IA impulsarán una cultura humana compartida y la búsqueda de respuestas más allá de cualquier cultura nacional o sistema de valores? ¿O amplificarán las plataformas de redes globales basadas en IA las lecciones o patrones específicos de los usuarios, con efectos diferentes a los que sus desarrolladores humanos planearon o anticiparon, o incluso socavarán tales efectos? No podemos evitar responder a estas preguntas porque nuestras comunicaciones, tal y como están construidas ahora, ya no pueden funcionar sin redes asistidas por IA.

Plataformas de redes y desinformación

Las fronteras nacionales han sido durante mucho tiempo permeables a nuevas ideas y tendencias, incluidas las fomentadas con un propósito deliberadamente maligno, pero nunca a tal escala. Aunque existe un amplio consenso sobre la importancia de evitar que la desinformación malévola distribuida intencionadamente impulse las

tendencias sociales y los acontecimientos políticos, garantizar este resultado rara vez ha resultado una empresa precisa o totalmente exitosa. Sin embargo, en el futuro, tanto la «ofensiva» como la «defensa» —tanto la difusión de la desinformación como los esfuerzos para combatirla— se automatizarán cada vez más y se confiarán a la IA. La IA generadora de lenguaje GPT-3 ha demostrado su capacidad para crear personajes ficticios, utilizarlos para producir un lenguaje característico del discurso del odio y entablar conversaciones con usuarios humanos para inculcarles prejuicios e incluso incitarles a la violencia.⁷ Si una IA de este tipo se desplegara para difundir el odio y la división a gran escala, los seres humanos por sí solos podrían no ser capaces de combatir el resultado. A menos que la IA se detenga en una fase temprana de su despliegue, identificar y desactivar manualmente todo su contenido mediante investigaciones y decisiones individuales resultaría muy difícil, incluso para los gobiernos y los operadores de plataformas de redes más sofisticados. Para una tarea tan vasta y ardua, tendrían que recurrir —como ya lo hacen— a algoritmos de IA de moderación de contenidos. Pero, ¿quién y cómo los crea y supervisa?

Cuando una sociedad libre confía en plataformas de redes asistidas por IA para generar, transmitir y filtrar contenidos a través de las fronteras nacionales y regionales, y cuando esas plataformas actúan de manera que promueven inadvertidamente el odio y la división, esa sociedad se enfrenta a una nueva amenaza que debería llevarla a considerar nuevos enfoques para vigilar su entorno de información. El problema subyacente es urgente, pero las soluciones basadas en la inteligencia artificial plantean sus propias cuestiones críticas. No debemos dejar de considerar el equilibrio adecuado entre el juicio humano y la automatización impulsada por la IA en ambos lados de la ecuación.

Para las sociedades acostumbradas al libre intercambio de ideas, lidiar con el papel de la IA en la evaluación y posible censura de la información ha introducido difíciles debates fundamentales. A

medida que las herramientas para difundir desinformación se vuelven más potentes y se automatizan cada vez más, el proceso de definir y suprimir la desinformación emerge con más fuerza como una función social y política esencial. Para las empresas privadas y los gobiernos democráticos, esta función supone no solo un grado de influencia y responsabilidad inusitado, sino también a menudo no buscado, sobre los cambios en los fenómenos sociales y culturales, que antes no eran operados o controlados por un único actor, sino que se habían desarrollado a través de millones de interacciones individuales en el mundo físico.

Para algunos, la inclinación será confiar la tarea a un proceso técnico que parezca libre de prejuicios y parcialidades humanas: una IA con una función objetivo para identificar y detener el flujo de desinformación y falsedad. Pero, ¿qué ocurre con los contenidos que nunca son vistos por el público? Cuando la prominencia o la difusión de un mensaje se ven tan limitadas que, de hecho, se niega su existencia, hemos llegado a un estado de censura. Si la IA antidesinformación comete un error, suprimiendo contenidos que no son desinformación maligna, sino sobre hecho auténticos, ¿cómo los identificamos? ¿Podemos saber lo suficiente, y a tiempo, para corregirlo? Por otra parte, ¿tenemos derecho a leer, o incluso un interés legítimo en leer, información «falsa» generada por la IA? El poder de entrenar a una IA defensiva contra una norma objetiva (o subjetiva) de falsedad —y la capacidad, si es que puede desarrollarse alguna, de supervisar las operaciones de esa IA— se convertirían en sí mismos en una función de importancia e influencia que rivalizaría con las funciones tradicionalmente desempeñadas por el gobierno. Pequeñas diferencias en el diseño de la función objetivo de una IA, los parámetros de entrenamiento y las definiciones de falsedad podrían provocar diferencias en los resultados capaces de alterar la sociedad. Estas cuestiones adquieren una importancia vital a medida que las plataformas de redes utilizan la IA para prestar sus servicios a miles de millones de personas.

Los debates políticos y normativos internacionales en torno a TikTok, una plataforma de redes basada en la inteligencia artificial para crear y compartir vídeos cortos y a menudo caprichosos, ofrecen una primera visión inesperada de los problemas que pueden surgir al confiar en la inteligencia artificial para dar forma a las comunicaciones, sobre todo cuando esa inteligencia artificial se desarrolla en una nación y la utilizan ciudadanos de otra. Los usuarios de TikTok graban y publican vídeos con sus teléfonos inteligentes, y muchos millones de usuarios disfrutan viéndolos. Algoritmos de IA recomiendan contenidos que podrían gustar a esas personas basándose en su uso previo de la plataforma. Desarrollada en China y popularizada en todo el mundo, TikTok no crea contenidos ni parece establecer grandes restricciones sobre ellos, más allá de un límite de tiempo para los vídeos y unas directrices comunitarias que prohíben la «desinformación», el «extremismo violento» y ciertos tipos de contenido gráfico.

Para el espectador general, el principal atributo de la visión del mundo asistida por IA de TikTok parece ser la veleidad: su contenido consiste principalmente en fragmentos de vídeos cortos y tontos de bailes, bromas y habilidades inusuales. Sin embargo, debido a la preocupación de los gobiernos por la recopilación de datos de los usuarios y su capacidad latente para la censura y la desinformación, tanto el gobierno indio como el estadounidense decidieron restringir el uso de TikTok en 2020. Además, Washington forzó la venta de las operaciones de TikTok en Estados Unidos a una empresa estadounidense que pudiera conservar los datos de los usuarios en el país, e impedir su exportación a China. A su vez, Pekín actuó para prohibir la exportación del código en el que se basa el algoritmo de recomendación de contenidos que constituye el núcleo de la eficacia de TikTok y de su atractivo para los usuarios.

Pronto, más plataformas de redes —quizás la mayoría de las que permiten la comunicación, el entretenimiento, el comercio, las finanzas y los procesos industriales— dependerán de una IA cada vez más sofisticada y hecha a medida para ofrecer funciones claves

y moderar y dar forma a los contenidos, a menudo más allá de las fronteras nacionales. Las ramificaciones políticas, jurídicas y tecnológicas de estas maniobras aún no han concluido. El hecho de que una sola aplicación de entretenimiento caprichoso con IA haya provocado tal consternación multinacional oficial sugiere que nos esperan enigmas geopolíticos y normativos más complejos en un futuro próximo.

Gobiernos y regiones

Las plataformas de redes plantean nuevos dilemas culturales y geopolíticos no solo para cada país, sino también, dada la natural ausencia de fronteras de esta tecnología, para las relaciones entre gobiernos y regiones más amplias. Incluso con una intervención gubernamental sustancial y sostenida, la mayoría de los países — incluso los tecnológicamente avanzados— no crearán empresas que produzcan o mantengan una versión «nacional» avanzada de cada plataforma de redes de influencia mundial (como las que se utilizan para los medios sociales, la búsqueda en la web, etc.). El ritmo del cambio tecnológico es demasiado rápido, y el número de programadores, ingenieros y profesionales del diseño y desarrollo de productos con conocimientos, demasiado escaso, para una cobertura tan amplia. La demanda mundial de talento es demasiado alta, los mercados locales de la mayoría de los servicios demasiado pequeños y los costes de los productos y servicios demasiado elevados para mantener una versión independiente de cada plataforma de redes. Para mantenerse en la vanguardia evolutiva del desarrollo tecnológico se requiere un capital intelectual y financiero superior al que poseen la mayoría de las empresas y al que la mayoría de los gobiernos están dispuestos o son capaces de proporcionar. Pero incluso en un escenario así, muchos usuarios, si se les da la opción, preferirían no estar limitados a una plataforma de redes que albergue solo a sus compatriotas y las ofertas de software y contenidos que producen. En su lugar, la dinámica de los efectos positivos de la red tenderá a apoyar solo a un puñado de

participantes que lideren la tecnología y el mercado de su producto o servicio concreto.

Muchos países dependen —y es probable que sigan dependiendo indefinidamente— de plataformas de redes diseñadas y alojadas en otros países. Por ende, también es probable que sigan dependiendo, al menos en parte, de los reguladores de otros países para el acceso continuo, los insumos clave y las actualizaciones internacionales. Por lo tanto, muchos gobiernos tendrán un incentivo para garantizar el funcionamiento continuado de los servicios en línea impulsados por la IA de otros países que ya se han incorporado a aspectos fundamentales de su sociedad. Este compromiso puede adoptar la forma de regular a los propietarios u operadores de las plataformas de redes, instituir requisitos para su funcionamiento o gestionar el entrenamiento de su IA. Los gobiernos podrían insistir en que los desarrolladores incluyan medidas para evitar ciertas formas de parcialidad o abordar determinados dilemas éticos.

Las figuras públicas pueden conseguir aprovechar una plataforma de redes y su IA para obtener una mayor visibilidad de sus contenidos, lo cual les permite llegar a audiencias más amplias. Pero si los operadores de la plataforma deciden que esas figuras prominentes han violado las normas de contenido, pueden ser fácilmente censuradas o eliminadas, impidiendo que puedan llegar a audiencias tan amplias (o llevando a sus audiencias a la clandestinidad). O sus contenidos podrían ir acompañados de algún tipo de etiqueta de advertencia u otra calificación potencialmente estigmatizadora. La cuestión es qué persona o institución debe tomar esa decisión. La autoridad para tomar y aplicar de forma independiente tales decisiones, que ahora recae en algunas empresas, refleja un nivel de poder que pocos gobiernos democráticos han ejercido. Aunque la mayoría de la gente consideraría indeseable que las empresas privadas tuvieran este grado de poder y control, cederlo a organismos gubernamentales sería casi igual de problemático; hemos ido más allá de los planteamientos políticos convencionales. En lo que respecta a las

plataformas de redes, la necesidad de este tipo de evaluaciones y decisiones ha surgido rápida y casi accidentalmente en los últimos años, y parece haber sorprendido a usuarios, gobiernos y empresas por igual. Es necesario resolverlo.

Plataformas de redes y geopolítica

La geopolítica emergente de las plataformas de redes constituye un nuevo aspecto clave de la estrategia internacional, y los gobiernos no son los únicos actores. Los gobiernos pueden tratar cada vez más de limitar el uso o el comportamiento de estos sistemas o intentar impedir que superen a sus rivales nacionales en regiones importantes, para evitar que una sociedad o economía competidora adquiera una poderosa influencia sobre el desarrollo industrial, económico o (más difícil de definir) político y cultural de ese país. Sin embargo, como los gobiernos no suelen crear ni gestionar estas plataformas de redes, las acciones de los inventores, las empresas y los usuarios individuales configurarán el campo junto con las restricciones o incentivos gubernamentales, en un escenario estratégico especialmente dinámico y difícil de predecir. Además, a esta ya compleja ecuación se añade una nueva forma de ansiedad cultural y política. En Pekín, Washington y algunas capitales europeas se ha expresado preocupación (y se ha articulado de forma oblicua en otros lugares) por las implicaciones de desarrollar amplios aspectos de la vida económica y social nacional en plataformas de redes facilitadas por IA diseñada en otros países, potencialmente rivales. A partir de este fermento tecnológico y político, se están estableciendo nuevas configuraciones geopolíticas.

Estados Unidos ha dado lugar a un conjunto de plataformas de redes privadas de alcance mundial y tecnológicamente punteras que dependen cada vez más de la IA. Las raíces de este logro se encuentran en el liderazgo académico de universidades que atraen a los mejores talentos mundiales, un ecosistema de *start-ups* que permite a los participantes llevar rápidamente las innovaciones a escala y obtener beneficios de sus desarrollos, y el apoyo

gubernamental a la I+D avanzada (a través de la National Science Foundation, DARPA y otras agencias). La prevalencia del inglés como lengua global, la creación de normas tecnológicas propias o influenciadas por Estados Unidos y la aparición de una importante base nacional de clientes particulares y corporativos constituyen un entorno favorable para los operadores de plataformas de redes estadounidenses. Algunos de estos operadores evitan la participación gubernamental y consideran que sus intereses son fundamentalmente no nacionales, mientras que otros han aceptado los contratos y programas gubernamentales. En el extranjero, cada vez se les trata más (a menudo sin distinción) como creaciones y representantes de Estados Unidos, aunque en muchos casos el papel del gobierno estadounidense se limitaba a no estorbarles.

Estados Unidos ha empezado a considerar las plataformas de redes como un aspecto de la estrategia internacional, restringiendo las actividades nacionales de algunas plataformas extranjeras y limitando la exportación de algunos programas y tecnologías que podrían facilitar el crecimiento de competidores extranjeros. Al mismo tiempo, los reguladores federales y estatales han identificado las principales plataformas de redes nacionales como objetivos de acciones antimonopolio. A corto plazo, al menos, este impulso simultáneo hacia la preeminencia estratégica y la multiplicidad nacional puede empujar el desarrollo estadounidense en direcciones contradictorias.

Del mismo modo, China ha apoyado el desarrollo de plataformas de redes que ya son de escala mundial, pero que, al mismo tiempo, están preparadas para expandirse aún más. Aunque el planteamiento regulador de Pekín ha fomentado una competencia feroz entre los agentes tecnológicos nacionales (con los mercados mundiales como objetivo último), ha excluido en gran medida (o ha impuesto ofertas muy ajustadas) a los homólogos no chinos dentro de las fronteras de China. En los últimos años, Pekín también ha tomado medidas para dar forma a las normas tecnológicas internacionales y prohibir la exportación de tecnologías sensibles

desarrolladas en el país. Las plataformas de redes chinas predominan en China y las regiones cercanas, y algunas son líderes en los mercados mundiales. Algunas plataformas de redes chinas disfrutan de ventajas integradas en las comunidades de la diáspora china (las comunidades de habla china de Estados Unidos y Europa, por ejemplo, siguen utilizando mucho las funciones financieras y de mensajería de WeChat), pero su atractivo no se limita a los consumidores chinos. Tras dominar el agitado mercado nacional chino, las preeminentes plataformas de redes del país y su tecnología de IA están en condiciones de competir en el mercado mundial.

En determinados mercados, como Estados Unidos y la India, los gobiernos se han vuelto cada vez más francos en su enfoque de las plataformas de redes chinas (y otras tecnologías digitales chinas) como extensiones potenciales o de facto de los objetivos políticos del gobierno chino. Aunque esto puede ser cierto en algunos casos, las dificultades de algunos operadores de plataformas de redes chinas sugieren que las relaciones de las empresas con el Partido Comunista Chino pueden ser complejas y variadas en la práctica. Es posible que los operadores de plataformas de redes chinas no reflejen automáticamente los intereses del partido o del Estado; es probable que la correlación dependa de las funciones concretas de las plataformas de redes y de la medida en que sus operadores comprendan y naveguen por las líneas rojas tácitas del Gobierno.

En términos más generales, aunque el este y el sudeste asiáticos, cuna de empresas de alcance mundial, producen tecnologías clave como semiconductores, servidores y electrónica de consumo, también son el hogar de plataformas de redes creadas localmente. En toda la región, las plataformas creadas en China y Estados Unidos tienen mayor o menor influencia entre los distintos segmentos de la población. En sus relaciones con las plataformas de redes, como en otros aspectos de la economía y la geopolítica, los países de la región han estado estrechamente vinculados al ecosistema tecnológico de origen estadounidense. Pero también hay

un uso sustancial de las plataformas de redes chinas, así como un compromiso más amplio con las empresas y la tecnología chinas, que los asiáticos orientales y sudorientales pueden considerar orgánicamente conectadas a su región e integradas en su propio éxito económico.

A diferencia de China y Estados Unidos, Europa aún no ha creado plataformas de redes globales propias ni cultivado el tipo de industria nacional de tecnología digital que ha apoyado el desarrollo de grandes plataformas en otros lugares. Aun así, Europa atrae la atención de los principales operadores de plataformas de redes por sus empresas y universidades punteras, su tradición de exploración de la Ilustración, que sentó las bases esenciales de la era informática, su considerable mercado y un aparato regulador formidable en su capacidad para innovar e imponer requisitos legales. Sin embargo, Europa sigue enfrentándose a desventajas para el escalado inicial de las nuevas plataformas de redes debido a su necesidad de atender a muchas lenguas y aparatos reguladores nacionales para llegar a su mercado combinado. En cambio, las plataformas de redes nacionales de Estados Unidos y China son capaces de operar a escala continental, lo que permite a sus empresas afrontar mejor la inversión necesaria para seguir creciendo en otras lenguas.

La Unión Europea (UE) ha centrado recientemente su atención reguladora en las condiciones de participación de los operadores de plataformas de redes en su mercado, incluido el uso de la IA por parte de estos operadores (y otras entidades). Como en otras cuestiones geopolíticas, Europa se enfrenta a la disyuntiva de actuar como aliada de uno u otro bando en cada una de las grandes esferas tecnológicas —dibujando su rumbo mediante el establecimiento de una relación especial— o como equilibradora entre bandos.

En este sentido, las preferencias de los Estados tradicionales de la UE y de los recién llegados de Europa Central y Oriental pueden diferir, por sus situaciones geopolíticas y económicas distintas. Hasta ahora, las potencias mundiales históricas como Francia y Alemania

han valorado la independencia y la libertad de maniobra en su política tecnológica. Sin embargo, los Estados europeos periféricos con experiencia reciente y directa de amenazas extranjeras —como los Estados postsoviéticos bálticos y centroeuropeos— han mostrado una mayor disposición a identificarse con una «tecnosfera» liderada por Estados Unidos.

La India, aunque todavía es una fuerza emergente en este ámbito, cuenta con un capital intelectual considerable, un entorno empresarial y académico relativamente favorable a la innovación y una gran reserva de talento tecnológico y de ingeniería que podría respaldar la creación de plataformas de redes líderes (como se ha demostrado recientemente con su industria autóctona de compras en línea). La población y la economía de la India tienen un tamaño tal que podrían sostener plataformas de redes potencialmente independientes sin recurrir a otros mercados. Asimismo, las plataformas de redes diseñadas en la India tienen potencial para popularizarse también en otros mercados. En décadas anteriores, gran parte del talento indio en software se desplegó en la industria de servicios de tecnologías de la información (TI) o en plataformas de redes no indias. Ahora, a medida que el país evalúa sus relaciones regionales y su relativa dependencia de la tecnología importada, puede optar por trazar un camino más independiente o asumir un papel principal dentro de un bloque internacional de naciones tecnológicamente compatibles.

Rusia, a pesar de una formidable tradición nacional en matemáticas y ciencias, ha producido hasta ahora pocos productos y servicios digitales atractivos para el consumidor más allá de sus propias fronteras. Sin embargo, sus formidables capacidades cibernéticas y su demostrada habilidad para penetrar las defensas y llevar a cabo operaciones a través de redes globales sugieren que Rusia debe contarse entre las importantes potencias tecnológicas del mundo. Tal vez como resultado de explotar las vulnerabilidades en línea de otros países, Rusia también ha fomentado el uso de ciertas plataformas de redes a escala nacional (como las de búsqueda, por

ejemplo, Yandex), aunque en su forma actual, estas tienen un atractivo relativamente limitado para los consumidores no rusos. En la actualidad, estas plataformas funcionan como recurso o alternativa a los proveedores dominantes, no como competidores económicos sustanciales.

Configurada principalmente por estos gobiernos y regiones, se está desarrollando una contienda multidisciplinar por la ventaja económica, la seguridad digital, la primacía tecnológica y los objetivos éticos y sociales, aunque hasta la fecha los principales actores no han identificado de forma coherente la naturaleza de la contienda ni las reglas del juego.

Uno de los enfoques ha consistido en tratar las plataformas de redes y su IA como una cuestión de regulación nacional. Desde este punto de vista, el principal reto de los gobiernos es impedir que las plataformas abusen de sus posiciones o eludan responsabilidades previamente establecidas o reguladas. Estos conceptos están evolucionando y son objeto de controversia, sobre todo entre Estados Unidos y la UE. Y debido a la forma en que los efectos de red positivos aumentan el valor para los usuarios con la escala, estas responsabilidades a menudo resultan difíciles de definir.

Otro enfoque se ha centrado en tratar la aparición y el funcionamiento de las plataformas de redes principalmente como una cuestión de estrategia internacional. Desde este punto de vista, la popularización de un operador extranjero dentro de un país introduce nuevos factores culturales, económicos y estratégicos. Existe la preocupación de que las plataformas de redes puedan fomentar, aunque sea de forma pasiva, un nivel de conexión e influencia que antes solo habría surgido de una estrecha alianza, especialmente con el uso de la IA como herramienta para aprender de los ciudadanos e influir en ellos. Si una plataforma de redes es útil y tiene éxito, llega a apoyar funciones comerciales e industriales más amplias y, en esta capacidad, puede llegar a ser nacionalmente indispensable. Al menos teóricamente, la amenaza de retirada de una plataforma de redes (o de sus insumos tecnológicos claves), ya

sea por parte de un gobierno o de una empresa, sirve como instrumento potencial de influencia, pero también como incentivo para hacerla prescindible. Esta hipotética capacidad de convertir en armas las plataformas de redes (u otras tecnologías) al retenerlas en caso de crisis, puede incitar a los gobiernos a adoptar nuevas formas de política y estrategia.

Para los países y regiones que no producen plataformas de redes propias, la elección para su futuro inmediato parece estar entre (1) limitar la dependencia de plataformas que podrían proporcionar una ventaja a un gobierno adversario; (2) permanecer vulnerables —por ejemplo, a la capacidad potencial de otro gobierno para acceder a datos sobre sus ciudadanos—; o (3) contrarrestar las amenazas potenciales. Un gobierno puede decidir que los riesgos de permitir que determinadas plataformas de redes extranjeras operen dentro de sus fronteras son inaceptables, o que deberían equilibrarse con la introducción de plataformas de redes rivales. Los gobiernos con recursos pueden optar por patrocinar a un participante nacional como rival: en muchos casos, sin embargo, esta opción requeriría una intervención sustancial y sostenida, y aún así correría el riesgo de fracasar. Es probable que los países avanzados intenten evitar depender de los productos de cualquier otro país para funciones clave (por ejemplo, redes sociales, comercio, viajes compartidos), sobre todo en ámbitos en los que existen varias plataformas de redes disponibles a nivel mundial.

El hecho de que las plataformas de redes dotadas de IA creadas por una sociedad puedan funcionar y evolucionar dentro de otra y convertirse en algo inextricable de la economía y el discurso político nacional de ese país supone un cambio fundamental con respecto a épocas anteriores. Antes, las fuentes de información y comunicación solían tener un alcance local y nacional, y no mantenían una capacidad de aprendizaje independiente. Hoy en día, las plataformas de redes de transporte creadas en un país pueden convertirse en las arterias y la sangre vital de otro país, a medida que la plataforma aprende qué consumidores necesitan determinados productos y

automatiza la logística del suministro. En efecto, estas plataformas de redes podrían convertirse en infraestructuras económicas críticas, y dar al país de origen influencia sobre cualquier país que dependa de ellas.

Por el contrario, cuando los gobiernos deciden limitar el alcance de la tecnología extranjera en sus economías, sus decisiones pueden obstaculizar la difusión de esa tecnología, o incluso su viabilidad comercial continuada. Los gobiernos pueden centrarse en prohibir el uso de plataformas de redes extranjeras que hayan sido identificadas como amenazas. Varios países han adoptado medidas de este tipo para los productos extranjeros en general y para las plataformas de redes en particular. Este enfoque regulador puede crear tensiones con la expectativa de la población de que debe ser libre de utilizar lo que mejor funcione. En las sociedades abiertas, estas prohibiciones también pueden plantear cuestiones difíciles y novedosas sobre el alcance adecuado de la regulación gubernamental.

Atrapados entre las medidas gubernamentales y la preocupación por su estatus global y su base de usuarios, los operadores de plataformas de redes deberán decidir hasta qué punto se convierten, de hecho, en un conglomerado de empresas nacionales y/o regionales, potencialmente en varias jurisdicciones distintas. Por el contrario, pueden decidir comportarse como empresas globales que persiguen de forma independiente sus valores, que pueden no alinearse claramente con las prioridades de ningún gobierno en particular.

Tanto en Occidente como en China han aumentado las valoraciones oficiales de la importancia de los productos y servicios digitales de la otra parte, incluidas las plataformas de redes asistidas por IA. Y fuera de estos países, los gobiernos y los usuarios pueden ver las principales plataformas de redes como una expresión de la cultura o los intereses estadounidenses o chinos. Los valores y principios organizativos de los operadores de plataformas de redes pueden reflejar los de la sociedad de la cual surgieron, pero en

Occidente, al menos, no hay ningún requisito de tal correspondencia. Las culturas empresariales occidentales suelen valorar más la autoexpresión y la universalidad que el interés nacional o la conformidad con las tradiciones establecidas.

Incluso cuando no se ha producido una «disociación tecnológica» entre países o regiones, las medidas gubernamentales están empezando a clasificar a las empresas en campos distintos que atienden a grupos específicos de usuarios dedicados a actividades concretas. Y a medida que la IA aprende y se adapta a bases de usuarios geográfica o nacionalmente distintas, puede a su vez influir de manera diferente en el comportamiento humano de las distintas regiones. De este modo, una industria basada en la premisa de la comunidad y la comunicación globales puede, con el tiempo, facilitar un proceso de regionalización, uniendo bloques de usuarios en realidades separadas, influidos por IA distintas que han evolucionado en direcciones diferentes. Con el tiempo, podrían desarrollarse esferas de normas tecnológicas regionales, con diversas plataformas de redes habilitadas por IA y las actividades o expresiones que apoyan evolucionando a lo largo de líneas paralelas pero totalmente distintas y con una comunicación e intercambio entre ellas cada vez más ajenos y difíciles.

El tira y afloja de individuos, empresas, reguladores y gobiernos nacionales que intentan dar forma y canalizar las plataformas de redes habilitadas por IA será cada vez más complejo y se desarrollará alternativamente como una contienda estratégica, una negociación comercial y un debate ético. Cuestiones que parecen urgentes pueden estar desfasadas para cuando los participantes oficiales pertinentes se hayan reunido para debatirlas. Para entonces, la plataforma de redes habilitada por IA puede haber aprendido o mostrado un nuevo comportamiento que convierta los términos originales de la discusión en obsoletos o insuficientes. Los creadores y operadores pueden llegar a comprender mejor los objetivos y límites de las plataformas de redes, pero es poco probable que intuyan de antemano las probables preocupaciones

gubernamentales o las objeciones filosóficas más generales. Se necesita urgentemente un diálogo entre estos sectores sobre las preocupaciones y los planteamientos básicos, que, en la medida de lo posible, debería tener lugar antes de que la IA se despliegue como parte de plataformas de redes a gran escala.

Las plataformas de redes dotadas de IA y nuestro futuro humano

La percepción y la experiencia humanas, filtradas por la razón, han definido durante mucho tiempo nuestra comprensión de la realidad. Esta comprensión ha tenido normalmente un alcance individual y local, y solo ha alcanzado una correspondencia más amplia para determinadas cuestiones y fenómenos esenciales; rara vez ha sido global o universal, salvo en el contexto distintivo de la religión. Ahora, la realidad cotidiana es accesible a escala mundial, a través de plataformas de redes que unen a un gran número de usuarios. Sin embargo, la mente humana individual ya no es la única —o quizá ni siquiera la principal— navegante de la realidad. Las plataformas de redes globales y continentales basadas en la IA se han unido a la mente humana en esta tarea, ayudándola y, en algunas áreas, quizás desplazándola.

Hay que definir nuevos conceptos de entendimiento y limitaciones entre regiones, gobiernos y operadores de plataformas de redes. La mente humana nunca ha funcionado de la manera que exige la era de Internet. Con sus complejos efectos sobre la defensa, la diplomacia, el comercio, la sanidad y el transporte, que generan dilemas estratégicos, tecnológicos y éticos demasiado complejos para que los aborde un único actor o disciplina, la llegada de las plataformas de redes con IA está planteando cuestiones que no deben considerarse de naturaleza exclusivamente nacional, partidista o tecnológica.

Los estrategas deben tener en cuenta las lecciones de épocas anteriores. No deben asumir que la victoria total es posible en cada

contienda comercial y tecnológica. Por el contrario, deben reconocer que prevalecer requiere una definición de éxito que una sociedad pueda mantener a lo largo del tiempo. Esto, a su vez, implica responder al tipo de preguntas que eludieron los líderes políticos y los planificadores estratégicos durante la Guerra Fría: ¿Qué margen de superioridad será necesario? ¿En qué momento la superioridad deja de tener sentido en términos de rendimiento? ¿Qué grado de inferioridad seguiría teniendo sentido en una crisis en la que cada bando utilizara al máximo sus capacidades?

Los operadores de plataformas de redes se enfrentarán a opciones que van más allá de las de servir a los clientes y lograr el éxito comercial. Hasta ahora, por lo general no se han visto obligados a definir una ética nacional o de servicio más allá del impulso orgánico por mejorar sus productos, aumentar su alcance y servir a los intereses de usuarios y accionistas. Sin embargo, a medida que vayan asumiendo papeles más amplios e influyentes, incluidas funciones que influyen (y a veces rivalizan) con las actividades de los gobiernos, se enfrentarán a retos mucho mayores. No solo deberán ayudar a definir la capacidad y los fines últimos de los reinos virtuales que han creado, sino también prestar cada vez más atención a cómo interactúan entre sí y con otros sectores de la sociedad.

¹ <https://investor.fb.com/investor-news/press-release-details/2021/Facebook-Reports-Fourth-Quarter-and-Full-Year-2020-Results/default.aspx>.

² Las estadísticas sobre expulsiones se publican trimestralmente. Véase <https://transparency.facebook.com/community-standards-enforcement>.

³ Véase Cade Metz, «AI Is Transforming Google Search. The Rest of the Web Is Next», en Wired, February 4, 2016. Desde entonces, los avances en IA para búsquedas han continuado. Algunos de los más recientes se describen en el blog de Google, The Keyword (véase Prabhakar Raghavan, «How AI Is Powering a More Helpful Google», October 15, 2020,

<https://blog.google/products/search/search-on/>), como la corrección ortográfica y la posibilidad de buscar frases o pasajes concretos, vídeos y resultados numéricos.

⁴ Los efectos positivos de la red pueden entenderse mejor en relación con las economías de escala. Con las economías de escala, un gran proveedor puede tener a menudo una ventaja de costes y, cuando ello conduce a precios más bajos, puede beneficiar a los clientes o usuarios individuales. Pero como los efectos positivos de red se refieren a la eficacia de un producto o servicio, no solo a su coste, suelen ser considerablemente más fuertes que las economías de escala.

⁵ Véase Kris McGuffie y Alex Newhouse, «The Radicalization Risks Posed by GPT-3 and Advanced Neural Language Models», Middlebury Institute of International Studies at Monterey, Center on Terrorism, Extremism, and Counterterrorism, September 9, 2020, https://www.middlebury.edu/institute/sites/www.middlebury.edu.institute/files/2020-09/gpt3-article.pdf?fbclid=IwAR0r0LroOYpt5wgr8EOpsIvGL2sEAi5H0PimcGIQcrpKFAG_HDDs3lBgqpU.

⁶ Carl von Clausewitz: *De la guerra*, traducción de Carlos Fortea. La Esfera de los libros, Madrid, 2007.

⁷ Esta dinámica se extiende más allá del ámbito puramente militar. Véase Kai-Fu Lee, *AI Superpowers: China, Silicon Valley, and the New World Order* (Boston y New York: Houghton Mifflin Harcourt, 2018); Michael Kanaan, *T-Minus AI: Humanity's Countdown to Artificial Intelligence and the New Pursuit of Global Power* (Dallas: BenBella Books, 2020).

Seguridad y orden mundial

Desde que se tiene constancia de la historia, la seguridad ha sido el objetivo mínimo de una sociedad organizada. Las culturas han diferido en sus valores, y las unidades políticas han diferido en sus intereses y aspiraciones, pero ninguna sociedad que no pudiera defenderse —sola o alineada con otras sociedades— ha perdurado.

En todas las épocas, las sociedades en busca de seguridad han tratado de convertir los avances tecnológicos en métodos cada vez más eficaces para vigilar las amenazas, lograr una preparación superior, ejercer influencia más allá de sus fronteras y, en caso de guerra, habilitar la fuerza para prevalecer. Para las primeras sociedades organizadas, los avances en metalurgia, arquitectura de fortificaciones, caballería y construcción naval fueron a menudo decisivos. A principios de la Edad Moderna, las innovaciones en armas de fuego y cañones, buques de guerra e instrumentos y técnicas de navegación desempeñaron un papel comparable. Reflexionando sobre esta eterna dinámica en su clásico de 1832, *Sobre la guerra*, el teórico militar prusiano Carl von Clausewitz señalaba: «La fuerza, para contrarrestar a la fuerza contraria, se equipa con los inventos del arte y la ciencia».¹

Algunas innovaciones, como la construcción de murallas y fosos, han favorecido la defensa. Sin embargo, a lo largo de los siglos se ha ido dando prioridad a la adquisición de medios para proyectar el poder a través de distancias cada vez más largas y con mayor rapidez y fuerza. En la época de la Guerra Civil estadounidense (1861-1865) y de la Guerra Franco-Prusiana (1870-1871), los

conflictos militares habían entrado en la era de la máquina, asumiendo cada vez más las características de la guerra total, como la producción industrializada de armas, la transmisión de órdenes por telégrafo y el transporte de tropas y material por ferrocarril a través de distancias continentales.

Con cada aumento de poder, las grandes potencias se han tomado la medida unas a otras, evaluando qué bando prevalecería en un conflicto, qué riesgos y pérdidas conllevaría esa victoria, qué los justificaría y cómo afectaría al resultado la entrada de otra potencia y su arsenal. Las capacidades, objetivos y estrategias de las distintas naciones se fijaban, al menos teóricamente, en un equilibrio, o una balanza de poder.

En el último siglo, la estrategia de calibración de medios y fines se ha desajustado. Las tecnologías empleadas para perseguir la seguridad se han multiplicado y se han vuelto más destructivas, mientras las estrategias para utilizarlas con el fin de alcanzar los objetivos definidos se han vuelto más elusivas. En nuestra era, la llegada de las capacidades cibernéticas y de IA están añadiendo nuevos y extraordinarios niveles de complejidad y abstracción a estos cálculos.

En este proceso, la Primera Guerra Mundial (1914-1918) supuso una disyuntiva. A principios del siglo XX, las principales potencias de Europa —con economías avanzadas, comunidades científicas e intelectuales pioneras y una confianza ilimitada en sus misiones globales— aprovecharon los avances tecnológicos de la Revolución Industrial para construir ejércitos modernos. Acumularon masas de tropas mediante el servicio militar obligatorio y material transportable por tren, así como ametralladoras y otras armas de fuego de carga rápida. Desarrollaron métodos de producción avanzados para reabastecer los arsenales a «velocidad de máquina», armas químicas (cuyo uso está prohibido desde entonces, prohibición que han aceptado la mayoría de los gobiernos, aunque no todos) y buques de guerra blindados y tanques rudimentarios. Diseñaron elaboradas estrategias basadas en la consecución de

ventajas mediante una rápida movilización y alianzas basadas en férreos compromisos entre aliados para movilizarse de forma concertada, rápida y completa, en caso de provocación. Cuando surgió una crisis sin importancia global inherente —el asesinato del heredero al trono de los Habsburgo por un nacionalista serbio— las grandes potencias de Europa siguieron estos planes hasta desembocar en un conflicto general. El resultado fue una catástrofe que destruyó una generación en pos de unos resultados no relacionados con ninguno de los objetivos bélicos originales de las partes. Tres imperios fueron testigos del colapso de sus instituciones. Incluso los vencedores quedaron agotados durante décadas y sufrieron una disminución permanente de sus papeles internacionales. La combinación de inflexibilidad diplomática, tecnología militar avanzada y planes de movilización a la desesperada había creado un círculo vicioso que hizo posible, pero también inevitable, la guerra mundial. Las bajas eran tan enormes que la necesidad de justificarlas hacía imposible el compromiso.

Desde aquel cataclismo, a pesar de toda la atención, disciplina y recursos dedicados a sus arsenales, las grandes potencias han magnificado los enigmas de la estrategia moderna. Al final de la Segunda Guerra Mundial y durante las primeras décadas de la Guerra Fría, las dos superpotencias compitieron por construir armas nucleares y sistemas de lanzamiento intercontinentales, cuya vasta capacidad destructiva resultó plausiblemente relacionable solo con los objetivos estratégicos más graves y totales. Al observar la primera prueba de armas nucleares en los desiertos de Nuevo México, el físico J. Robert Oppenheimer, uno de los padres de la bomba atómica, se sintió movido a invocar no las máximas estratégicas de Clausewitz, sino una línea de la escritura hindú, el Bhagavad Gita: «Ahora me he convertido en la Muerte, la destructora de mundos». Esta intuición presagiaba la paradoja central de la estrategia de la Guerra Fría: la tecnología armamentística dominante de la época nunca se utilizó. La

destruictividad de las armas seguía siendo desproporcionada con respecto a objetivos alcanzables más allá de la pura supervivencia.

El vínculo entre capacidades y objetivos permaneció roto durante toda la Guerra Fría, o al menos no se conectó de forma que permitiera un desarrollo claro de la estrategia. Las grandes potencias construyeron ejércitos tecnológicamente avanzados y sistemas de alianzas, tanto regionales como globales, pero no los utilizaron entre sí ni en conflictos con países más pequeños o movimientos armados con arsenales más rudimentarios, una amarga verdad experimentada por Francia en Argelia, Estados Unidos en Corea, y Estados Unidos y la Unión Soviética en Afganistán.

La era de la ciberguerra y la IA

Hoy, tras la Guerra Fría, las grandes potencias y otros Estados han aumentado su arsenal con capacidades cibernéticas cuya utilidad deriva en gran medida de su opacidad y negación y, en algunos casos, de su funcionamiento, en la ambigua frontera de la desinformación, la recogida de información, el sabotaje y el conflicto tradicional, con estrategias sin doctrinas reconocidas. Mientras tanto, cada avance ha ido acompañado de nuevas vulnerabilidades.

La era de la IA corre el riesgo de complicar los enigmas de la estrategia moderna más allá de la intención humana, o quizá de toda comprensión humana. Incluso si las naciones se abstienen de desplegar de forma generalizada las llamadas armas autónomas letales —armas de IA automáticas o semiautomáticas entrenadas y autorizadas para seleccionar sus propios objetivos y atacar sin más autorización humana—, la IA puede aumentar las capacidades convencionales, nucleares y cibernéticas de forma que las relaciones de seguridad entre rivales sean más difíciles de predecir y mantener y los conflictos más difíciles de limitar.

Las posibles funciones defensivas de la IA operan a varios niveles y pronto podrían resultar indispensables. Los aviones de combate pilotados por IA ya han demostrado una gran capacidad para dominar a los pilotos humanos en combates simulados. Utilizando

algunos de los mismos principios generales que permitieron las victorias de AlphaZero y el descubrimiento de la halicina, la IA puede identificar patrones de conducta que un adversario no planificó o advirtió, y luego recomendar métodos para contrarrestarlos. La IA puede permitir la traducción simultánea o la transmisión instantánea de otra información crítica al personal en zonas de crisis, cuya capacidad para entender su entorno o hacerse entender puede ser esencial para una misión o para la seguridad personal.

Ningún país importante puede permitirse ignorar las dimensiones de seguridad de la IA. Ya se está produciendo una carrera por la ventaja estratégica de la IA, sobre todo entre Estados Unidos y China y, en cierta medida, Rusia.² A medida que se extienda el conocimiento —o la sospecha— de que otros están obteniendo ciertas capacidades de IA, más naciones las buscarán. Una vez introducidas, estas capacidades podrían extenderse rápidamente. Aunque la creación de una IA sofisticada requiere una potencia informática considerable, la proliferación o el funcionamiento de la IA no suele requerirla.

La solución a estas complejidades no es ni la desesperación ni el desarme. Existen tecnologías nucleares, cibernéticas y de IA. Cada una de ellas desempeñará inevitablemente un papel en la estrategia. Ninguna quedará «sin inventar». Si Estados Unidos y sus aliados retroceden ante las implicaciones de estas capacidades y detienen el progreso en ellas, el resultado no sería un mundo más pacífico. Por el contrario, sería un mundo menos equilibrado en el que el desarrollo y el uso de las capacidades estratégicas más formidables tendrían menos en cuenta los conceptos de responsabilidad democrática y equilibrio internacional. Tanto el interés nacional como el imperativo moral aconsejan que Estados Unidos no ceda estos campos; de hecho, Estados Unidos debería esforzarse por configurarlos.

El progreso y la competencia en estos campos implicarán transformaciones que pondrán a prueba los conceptos tradicionales de seguridad. Antes de que estas transformaciones alcancen un

punto de inexorabilidad, hay que hacer algún esfuerzo para definir las doctrinas estratégicas relacionadas con la IA y compararlas con las de otras potencias (tanto Estados como actores no estatales). En las próximas décadas, deberemos lograr un equilibrio de poder que tenga en cuenta los intangibles de los conflictos cibernéticos y la desinformación a gran escala, así como las cualidades distintivas de la guerra asistida por la IA. El realismo obliga a reconocer que los rivales de la IA, incluso mientras compiten, deberían esforzarse por explorar el establecimiento de límites al desarrollo y uso de capacidades de IA excepcionalmente destructivas, desestabilizadoras e impredecibles. Un esfuerzo sobrio por controlar las armas de IA no está reñido con la seguridad nacional; es un intento de garantizar que la seguridad se persigue y se logra en el contexto de un futuro humano.

Armas nucleares y disuasión

En épocas anteriores, cuando surgía una nueva arma, los ejércitos la integraban en sus arsenales y los estrategas diseñaban doctrinas que permitían su uso en pos de fines políticos. La aparición de las armas nucleares rompió este vínculo. El primer, y único hasta la fecha, uso de armas nucleares en una guerra —por parte de Estados Unidos contra Hiroshima y Nagasaki en 1945, que puso fin rápidamente a la Segunda Guerra Mundial en el Pacífico— fue reconocido inmediatamente como un hito. Mientras las principales potencias del mundo redoblaban sus esfuerzos para dominar la nueva tecnología armamentística e incorporarla a sus arsenales, entablaron un debate inusualmente abierto y profundo sobre las implicaciones estratégicas y morales de su uso.

Con un poder a una escala muy superior al de cualquier otra forma de armamento de la época, las armas nucleares planteaban cuestiones fundamentales: ¿Podría relacionarse esta tremenda fuerza destructiva, mediante algún principio rector o doctrina, con los elementos tradicionales de la estrategia? ¿Podría conciliarse el uso de armas nucleares con objetivos políticos que no fueran la

guerra total y la destrucción mutua? ¿Admitiría la bomba un uso calibrado, proporcional o táctico?

La respuesta, hasta la fecha, ha oscilado entre ambigua y negativa. Incluso durante el breve periodo en el que Estados Unidos tuvo el monopolio nuclear (de 1945 a 1949) —y en el periodo algo más largo durante el que poseyó sistemas de lanzamiento nuclear sustancialmente más eficaces— nunca desarrolló una doctrina estratégica ni identificó un principio moral que le persuadiera de utilizar armas nucleares en un conflicto real tras la Segunda Guerra Mundial. Después de eso, en ausencia de unas líneas doctrinales claras acordadas mutuamente por las potencias nucleares existentes —y quizás ni siquiera entonces— ningún responsable político podría saber qué seguiría a un uso «limitado» y si seguiría siendo limitado. Hasta la fecha, no se ha hecho tal intento. Durante una crisis en 1955 por los bombardeos a través del estrecho de Taiwán, el presidente Eisenhower —amenazando a la entonces no nuclear República Popular China si no se calmaba— comentó que no veía ninguna razón por la que las armas nucleares tácticas no pudieran usarse «exactamente igual que se usaría una bala o cualquier otra cosa».³ Casi siete décadas después, ningún líder ha puesto aún a prueba esta proposición.

En cambio, durante la Guerra Fría, el objetivo primordial de la estrategia nuclear pasó a ser la disuasión: el uso de armas, principalmente mediante la voluntad declarada de desplegarlas, para evitar que un adversario entrara en acción, ya fuera iniciando un conflicto o utilizando sus propias armas nucleares en uno. En esencia, la disuasión nuclear era una estrategia psicológica de objetivos negativos. Pretendía persuadir a un adversario para que no actuara mediante la amenaza de un contraataque. Esta dinámica dependía tanto de las capacidades físicas de un Estado como de una cualidad intangible: el estado de ánimo del agresor potencial y la capacidad de su oponente para moldearlo. Desde el punto de vista de la disuasión, una debilidad aparente podía tener las mismas consecuencias que una deficiencia real; un farol tomado en serio

podía resultar un elemento disuasorio más útil que una amenaza de buena fe ignorada. Única entre las estrategias de seguridad (al menos hasta ahora), la disuasión nuclear se basa en una serie de abstracciones imposibles de probar: la potencia disuasoria no puede demostrar cómo o por qué margen se ha evitado algo.

A pesar de estas paradojas, los arsenales nucleares se incorporaron a los conceptos básicos del orden internacional. Cuando Estados Unidos poseía el monopolio nuclear, su arsenal se utilizaba para disuadir los ataques convencionales y extender un «paraguas nuclear» sobre los países libres o aliados. Un avance soviético a través de Europa Occidental se veía frenado por la perspectiva, por remota o abstracta que fuera, de que Estados Unidos utilizaría armas nucleares para detener el ataque. Una vez que la Unión Soviética cruzó el umbral nuclear, el principal objetivo de las armas nucleares de ambas superpotencias pasó a ser cada vez más disuadir del uso de esas armas al otro bando. Se confiaba en la existencia de capacidades nucleares «supervivientes» —es decir, armas nucleares que pudieran lanzarse en un contraataque tras un hipotético primer ataque del adversario— para disuadir de la propia guerra nuclear. Y logró ese objetivo con respecto al conflicto entre las superpotencias.

Las hegemonías de la Guerra Fría gastaron enormes recursos en ampliar sus capacidades nucleares, al mismo tiempo que sus arsenales se alejaban cada vez más de la conducción cotidiana de la estrategia. La posesión de estos arsenales no disuadió a los Estados no nucleares —China, Vietnam, Afganistán— de desafiar a las superpotencias, ni impidió que los países de Europa Central y Oriental exigieran autonomía frente a Moscú.

Durante la guerra de Corea, la Unión Soviética era la única potencia nuclear aparte de Estados Unidos, y este último poseía una ventaja decisiva en número de armas y medios de lanzamiento. Sin embargo, los responsables políticos estadounidenses se abstuvieron de utilizarlas, optando por sufrir decenas de miles de bajas en batallas al estilo de la Primera Guerra Mundial contra fuerzas chinas

y norcoreanas no nucleares alineadas con la Unión Soviética (en retrospectiva, tenuemente) en lugar de aceptar la incertidumbre o el oprobio moral de una escalada nuclear. Desde entonces, todas las potencias nucleares que se han enfrentado a un adversario no nuclear han llegado a la misma conclusión, incluso cuando se han enfrentado a la derrota a manos de su enemigo no nuclear.

Durante esta época, a los responsables políticos no les faltaban estrategias. Bajo la doctrina de represalias masivas de los años 50, Estados Unidos amenazaba con responder a cualquier ataque, nuclear o convencional, con una escalada nuclear masiva. Sin embargo, una doctrina diseñada para convertir cualquier conflicto, por pequeño que fuera, en el Armagedón, resultó psicológica y diplomáticamente insostenible, además de parcialmente ineficaz. En respuesta, algunos estrategas propusieron doctrinas que permitirían el uso de armas nucleares tácticas en una guerra nuclear limitada.⁴ Pero estas propuestas fracasaron debido a las preocupaciones sobre la escalada y los límites. Los responsables políticos temían que las líneas doctrinales propuestas por los estrategas fueran demasiado ilusorias para detener la escalada hacia una guerra nuclear global. En consecuencia, la estrategia nuclear siguió centrándose en la disuasión y en garantizar la credibilidad de las amenazas, incluso en condiciones apocalípticas superiores a las que ningún ser humano había experimentado jamás durante una guerra. Estados Unidos distribuyó sus armas geográficamente y construyó una tríada (tierra, mar y aire) de capacidades de lanzamiento, asegurando que incluso un primer ataque sorpresa por parte de un adversario no impediría a Estados Unidos montar una respuesta devastadora.⁵ Se dice que los soviéticos exploraron el uso de un sistema diseñado para ser capaz, una vez encendido por usuarios humanos, de detectar un ataque nuclear entrante y difundir órdenes de lanzamiento para un contraataque sin más intervención humana; una exploración temprana del concepto de guerra semiautomatizada que implica la delegación de ciertas funciones de mando a una máquina.⁶

Los estrategas del gobierno y del mundo académico consideraron inquietante la dependencia de los ataques nucleares sin una contrapartida defensiva. Exploraron sistemas defensivos que, al menos en teoría, ampliarían el margen de decisión de los responsables políticos durante un enfrentamiento nuclear, y darían una oportunidad para la diplomacia o, como mínimo, para recabar más información y corregir interpretaciones erróneas. Irónicamente, sin embargo, la búsqueda de sistemas defensivos no hizo sino acelerar la demanda de armas ofensivas para penetrar las defensas de ambos bandos.

A medida que crecían los arsenales de ambas superpotencias, la posibilidad de desplegar realmente armas nucleares para prevenir o castigar las acciones de la otra parte llegó a parecer cada vez más surrealista e increíble, y a amenazar potencialmente la propia lógica de la disuasión. El reconocimiento de este punto muerto nuclear produjo una nueva doctrina con un nombre a medio camino entre la amenaza y el reconocimiento sardónico: destrucción mutua asegurada o MAD. Como el número de bajas asumido por esta teoría, que reducía los objetivos a la vez que agravaba la destructividad, era enorme, las armas nucleares se limitaron cada vez más al ámbito de la señalización y al aumento de la preparación de sistemas y unidades claves, avanzando gradualmente hacia los preparativos para un lanzamiento nuclear, en formas que debían ser advertidas y atendidas. Pero incluso el envío de tales señales se hacía con moderación, no fuera a ser que los adversarios las malinterpretaran y desencadenaran una catástrofe mundial. En busca de la seguridad, la humanidad había producido un arma definitiva y elaboradas doctrinas estratégicas para acompañarla. El resultado fue una ansiedad generalizada ante la posibilidad de que ese armamento llegara a utilizarse. El control de armamentos era un concepto destinado a apaciguar este dilema.

Control de armamento

Mientras la disuasión pretendía evitar la guerra nuclear mediante la amenaza, el control de armamentos pretendía evitarla mediante la limitación o incluso la abolición de las propias armas (o categorías de armas). Este enfoque se combinó con la no proliferación: el concepto, sustentado por un elaborado conjunto de tratados, salvaguardias técnicas y mecanismos reguladores y de control de otro tipo, de que las armas nucleares y el conocimiento y la tecnología que apoyan su construcción deben evitarse para que no se extiendan más allá de las naciones que ya las poseen. Ni el control de armamentos ni las medidas de no proliferación se habían intentado a tal escala para ninguna tecnología armamentística anterior. Hasta la fecha, ninguna de estas estrategias ha tenido éxito. Tampoco se ha aplicado en serio ninguna de las nuevas clases de armas, cibernéticas y de IA, inventadas en la era posterior a la Guerra Fría. Sin embargo, a medida que se multiplican los participantes en los ámbitos nuclear, cibernético y de la IA, la era del control de armamentos sigue ofreciendo lecciones dignas de consideración.

Tras las tensiones nucleares y el conflicto aparentemente cercano de la Crisis de los Misiles de Cuba (en octubre de 1962), las dos superpotencias de entonces, Estados Unidos y la Unión Soviética, intentaron circunscribir la competición nuclear a través de la diplomacia. Incluso mientras sus arsenales crecían y los de China, Gran Bretaña y Francia entraban en un cálculo de disuasión, Washington y Moscú autorizaron a sus negociadores a entablar un diálogo más sustantivo sobre el control de armamentos. Con cautela, buscaron límites en el número de armas y capacidades nucleares compatibles con el mantenimiento del equilibrio estratégico. Finalmente, ambas partes acordaron limitar no solo sus arsenales ofensivos sino también —siguiendo la paradójica lógica de la disuasión, en la que la vulnerabilidad se consideraba garantía de paz— sus capacidades defensivas. El resultado fue el Tratado de Limitación de Armas Estratégicas y el Tratado de Misiles Antibalísticos, de la década de 1970, y finalmente el Tratado de

Reducción de Armas Estratégicas (START), de 1991. En todos los casos, los límites impuestos a las armas ofensivas preservaron la capacidad de las superpotencias para destruirse mutuamente —y, por tanto, presumiblemente para disuadirse—, al tiempo que moderaron las carreras armamentísticas inspiradas en estrategias de disuasión.

Aunque siguieron siendo adversarios y continuaron disputándose ventajas estratégicas, Washington y Moscú consiguieron cierta seguridad en sus cálculos gracias a las negociaciones sobre el control de armamentos. Al informarse mutuamente sobre sus capacidades estratégicas y acordar ciertos límites básicos y mecanismos de verificación, ambos trataron de hacer frente al temor de que el otro aprovechara repentinamente una ventaja en una clase de armas nucleares para golpear primero.

En última instancia, estas iniciativas fueron más allá del objetivo de la autocontención y desalentaron activamente una mayor proliferación. A mediados de la década de 1960, Estados Unidos y Rusia crearon un régimen de múltiples compromisos y mecanismos destinado a prohibir la adquisición o posesión de armas nucleares a todos los Estados nucleares excepto a los originales, a cambio de compromisos para ayudar a otros Estados a aprovechar la tecnología nuclear para las energías renovables. Estos resultados se vieron facilitados por un sentimiento compartido distintivo sobre las armas nucleares —en la política, la cultura y en las relaciones entre los líderes individuales de la Guerra Fría— que reconocía que una guerra nuclear entre grandes potencias implicaría decisiones irreversibles y riesgos únicos para vencedores, vencidos y espectadores por igual.

Las armas nucleares plantearon a los responsables políticos dos enigmas persistentemente relacionados: cómo definir la superioridad y cómo limitar la inferioridad. En una época en la que las dos superpotencias poseían armamento suficiente para destruir el mundo varias veces, ¿qué significaba superioridad? Una vez que se había construido y desplegado un arsenal de forma creíblemente capaz de sobrevivir, el vínculo entre la adquisición de armas

adicionales, las ventajas obtenidas y los objetivos servidos se volvía opaco. Al mismo tiempo, un puñado de naciones adquirieron sus propios arsenales nucleares modestos, calculando que solo necesitaban un arsenal suficiente para infligir devastación —no para lograr la victoria— con el fin de disuadir ataques.

La no utilización nuclear no es un logro inherentemente permanente. Es una condición que debe ser asegurada por cada generación sucesiva de líderes para ajustar los despliegues y capacidades de sus armas más destructivas a una tecnología que evoluciona a una velocidad sin precedentes. Esto será especialmente difícil a medida que nuevos participantes con distintas doctrinas estratégicas y distintas actitudes hacia la provocación deliberada de víctimas civiles intenten desarrollar capacidades nucleares y a medida que las ecuaciones de disuasión se vuelvan cada vez más difusas e inciertas. En este mundo de paradojas estratégicas sin resolver están surgiendo nuevas capacidades, con las complejidades que estas conllevan.

El primero es el conflicto cibernético, que ha magnificado las vulnerabilidades, así como ampliado el campo de las contiendas estratégicas y la variedad de opciones a disposición de los participantes. El segundo es la IA, capaz de transformar la estrategia de las armas convencionales, nucleares y cibernéticas. La aparición de nuevas tecnologías ha agravado los dilemas de las armas nucleares.

Conflictos en la era digital

A lo largo de la historia, la influencia política de una nación ha tendido a ser aproximadamente correlativa a su poder militar y a sus capacidades estratégicas: su capacidad, aunque se ejerza principalmente mediante amenazas implícitas, para infligir daños a otras sociedades. Sin embargo, un equilibrio basado en un cálculo del poder no es estático ni se mantiene por sí mismo, sino que depende en primer lugar de un consenso sobre los elementos constitutivos del poder y los límites legítimos de su uso. Asimismo, el

mantenimiento del equilibrio requiere evaluaciones congruentes entre todos los miembros del sistema —especialmente los rivales— sobre las capacidades e intenciones relativas de los Estados, así como sobre las consecuencias de la agresión. Por último, la preservación del equilibrio requiere un equilibrio real y reconocido. Cuando un participante en el sistema aumenta desproporcionadamente su poder sobre los demás, el sistema intentará ajustarse, ya sea mediante la organización de una fuerza compensatoria o la adaptación a una nueva realidad. Cuando el cálculo del equilibrio se vuelve incierto, o cuando las naciones llegan a cálculos de poder relativo fundamentalmente diferentes, el riesgo de conflicto por error de cálculo alcanza su punto álgido.

En nuestra era, estos cálculos han entrado en un nuevo reino de abstracción. Esta transformación incluye las llamadas armas cibernéticas, una clase de armas que implican capacidades civiles de doble uso, de modo que su estatus como armas es ambiguo. En algunos casos, su utilidad para ejercer y aumentar el poder se debe en gran medida a que sus usuarios no revelan su existencia ni reconocen toda su gama de capacidades. Tradicionalmente, a las partes en conflicto no les costaba reconocer que se había producido un enfrentamiento ni saber quiénes eran los beligerantes. Los adversarios calculaban las capacidades de sus rivales y evaluaban la rapidez con la que podían desplegar sus arsenales. Ninguna de estas verdades tradicionales se traslada directamente al ámbito cibernético.

Las armas convencionales y nucleares existen en el espacio físico, donde sus despliegues pueden percibirse y sus capacidades calcularse al menos de forma aproximada. En cambio, las armas cibernéticas derivan una parte importante de su utilidad de su opacidad; su revelación puede degradar efectivamente algunas de sus capacidades. Sus intrusiones aprovechan fallos no revelados previamente en el software, y obtienen acceso a una red o sistema sin el permiso o conocimiento del usuario autorizado. En la contingencia de ataques distribuidos de denegación de servicio

(DDoS) (como en los sistemas de comunicación), se puede utilizar un enjambre de peticiones de información aparentemente válidas para saturar los sistemas y hacer que no estén disponibles para su uso previsto. En tales casos, las verdaderas fuentes del ataque pueden quedar enmascaradas, lo cual hace difícil o imposible determinar (al menos en el momento) quién está atacando. Incluso uno de los casos más famosos de sabotaje industrial cibernético —la interrupción por Stuxnet de los ordenadores de control de fabricación utilizados en los esfuerzos nucleares iraníes— no ha sido reconocida formalmente por ningún gobierno.

Las armas convencionales y nucleares se pueden dirigir con relativa precisión, y los imperativos morales y legales exigen que se dirijan contra fuerzas e instalaciones militares. Las armas cibernéticas pueden afectar ampliamente a los sistemas informáticos y de comunicaciones, y a menudo alcanzan con especial fuerza a los sistemas civiles. Las ciberarmas también pueden ser cooptadas, modificadas y redistribuidas por otros actores para otros fines. En ciertos aspectos, esto hace que las ciberarmas sean similares a las armas biológicas y químicas, cuyos efectos pueden propagarse de forma imprevista y desconocida. En muchos casos, afectan a amplias capas de la sociedad, no solo a objetivos específicos en un campo de batalla.⁷

Los atributos que confieren a las armas cibernéticas su utilidad hacen que el concepto de control de armas cibernéticas sea difícil de conceptualizar o perseguir. Los negociadores del control de armas nucleares fueron capaces de revelar o describir una clase de ojivas sin negar la función de esa arma. Los negociadores del control de armas cibernéticas (todavía inexistentes) deberán resolver la paradoja de que el debate sobre la capacidad de un arma cibernética puede ser lo mismo que su confiscación (permitir al adversario parchear una vulnerabilidad) o su proliferación (permitir al adversario copiar el código o el método de intrusión).

Estos retos se complican aún más por la ambigüedad que rodea a los términos y conceptos cibernéticos claves. Varias formas de

intrusiones cibernéticas, propaganda en línea y guerra de la información son denominadas, por diversos observadores en distintos contextos, «ciberguerra», «ciberataques» y, en algunos comentarios, «acto de guerra». Pero este vocabulario es incierto y a veces se utiliza de forma incoherente. Algunas actividades, como las intrusiones en redes para recopilar información, pueden ser análogas a la recopilación tradicional de inteligencia, aunque a nuevas escalas. Otros ataques —como las campañas de injerencia electoral en las redes sociales emprendidas por Rusia y otras potencias— son un tipo de propaganda digitalizada, desinformación e intromisión política con un alcance y un impacto mayores que en épocas anteriores. Son posibles gracias a la expansión de la tecnología digital y las plataformas de redes en las que se desarrollan estas campañas. Otras acciones cibernéticas tienen la capacidad de infligir impactos físicos similares a los sufridos durante las hostilidades tradicionales. La incertidumbre sobre la naturaleza, el alcance o la atribución de una ciberacción puede hacer que factores aparentemente básicos se conviertan en objeto de debate: por ejemplo, si se ha iniciado un conflicto, con quién o con qué se entabla el conflicto y hasta qué punto de la escalada puede llegar el conflicto entre las partes. En ese sentido, los principales países están inmersos en una especie de conflicto cibernético, aunque sin una naturaleza o alcance fácilmente definibles.⁸

Una paradoja central de nuestra era digital es que, cuanto mayor es la capacidad digital de una sociedad, más vulnerable se vuelve. Los ordenadores, los sistemas de comunicaciones, los mercados financieros, las redes eléctricas (y los sistemas digitales de mando y control de los que dependen) —incluso la mecánica de la política democrática— implican sistemas que son, en diversos grados, vulnerables a la manipulación o el ataque cibernéticos. A medida que las economías avanzadas integran sistemas digitales de mando y control en centrales eléctricas y redes eléctricas, trasladan sus programas gubernamentales a grandes servidores y sistemas en la nube y transfieren datos a libros de contabilidad electrónicos, su

vulnerabilidad a los ciberataques se multiplica; presentan un conjunto más rico de objetivos, de modo que un ataque exitoso podría ser sustancialmente devastador. Por el contrario, en el caso de una interrupción digital, el Estado de baja tecnología, la organización terrorista, e incluso los atacantes individuales pueden evaluar que tienen relativamente mucho menos que perder.

El coste comparativamente bajo de las capacidades y operaciones cibernéticas, y la relativa posibilidad de negación de su autoría, ha animado a algunos Estados a utilizar actores semiautónomos para realizar funciones cibernéticas. Al igual que los grupos paramilitares que invadieron los Balcanes en vísperas de la Primera Guerra Mundial, estos atacantes pueden ser difíciles de controlar y pueden llevar a cabo actividades provocadoras sin sanción oficial. Además de los filtradores y sabotadores capaces de neutralizar partes significativas de la capacidad cibernética de un Estado o agitar su panorama político nacional (incluso si estas actividades no escalan al nivel de un conflicto armado tradicional), la velocidad y la imprevisibilidad del dominio cibernético y la variedad de sus actores pueden tentar a los responsables políticos a tomar medidas preventivas para evitar un golpe de gracia.⁹

La velocidad y la ambigüedad del ámbito cibernético han favorecido la ofensiva y han fomentado los conceptos de «defensa activa» y «defensa anticipada», que tratan de interrumpir e impedir los ataques.¹⁰ El grado en que es posible la disuasión cibernética depende en parte de lo que un defensor pretenda disuadir y de cómo se mida el éxito. Los ataques más eficaces suelen ser los que se producen (a menudo sin reconocimiento inmediato o reconocimiento formal) por debajo del umbral de las definiciones tradicionales de conflicto armado. Ningún actor cibernético importante, gubernamental o no gubernamental, ha revelado toda la gama de sus capacidades o actividades, ni siquiera para disuadir las acciones de otros. La estrategia y la doctrina evolucionan con incertidumbre en un ámbito en la sombra, incluso cuando surgen nuevas capacidades. Estamos al principio de una frontera estratégica

que requiere una exploración sistémica, una estrecha colaboración entre el gobierno y la industria para garantizar unas capacidades de seguridad competitivas y —a su debido tiempo y con las salvaguardias adecuadas— un debate entre las grandes potencias sobre los límites.

La IA y la alteración en la seguridad

A la destructividad de las armas nucleares y los misterios de las armas cibernéticas se unen cada vez más nuevas clases de capacidades basadas en los principios de la inteligencia artificial analizados en capítulos anteriores. En silencio, a veces tímidamente, pero con un impulso inequívoco, las naciones están desarrollando y desplegando IA que facilita la acción estratégica a través de una amplia gama de capacidades militares, con efectos potencialmente revolucionarios en la política de seguridad.^{[11](#)}

La introducción de la lógica no humana en los sistemas y procesos militares transformará la estrategia. Los ejércitos y los servicios de seguridad que se entrenen o se asocien con la IA lograrán percepciones e influencia que sorprenderán y, en ocasiones, inquietarán. Estas asociaciones pueden negar o reforzar decisivamente aspectos de las estrategias y tácticas tradicionales. Si se delega en la IA cierto grado de control sobre las armas cibernéticas (ofensivas o defensivas) o físicas, como las aeronaves, puede llevar a cabo rápidamente funciones que los humanos realizan con dificultad. Inteligencias artificiales como μ Zero, de las Fuerzas Aéreas estadounidenses, ya han pilotado aviones y manejado sistemas de radar durante vuelos de prueba. En el caso de μ Zero, sus desarrolladores la diseñaron para tomar «decisiones finales» sin intervención humana, pero limitaron sus capacidades a pilotar un avión y manejar un sistema de radar.^{[12](#)} Es posible que otros países y equipos de diseño actúen con menos moderación.

Además de su utilidad potencialmente transformadora, la capacidad de autonomía y lógica independiente de la IA genera una

capa de *incalculabilidad*. La mayoría de las estrategias y tácticas militares tradicionales se han basado en el supuesto de un adversario humano cuya conducta y cálculo de toma de decisiones se ajustan a un marco reconocible o han sido definidos por la experiencia y la sabiduría convencional. Sin embargo, una IA que pilote un avión o busque objetivos sigue su propia lógica, que puede ser inescrutable para un adversario e insensible a las señales y fintas tradicionales, y que, en la mayoría de los casos, avanzará más rápido que la velocidad del pensamiento humano.

La guerra siempre ha sido un ámbito de incertidumbre y contingencia, pero la entrada de la IA en este espacio introducirá nuevas dimensiones. Como las IA son dinámicas y emergentes, ni siquiera las potencias que crean o manejan un arma diseñada u operada por una IA saben exactamente lo poderosa que es o lo que hará en una situación determinada. ¿Cómo se desarrolla una estrategia —ofensiva o defensiva— para algo que percibe aspectos del entorno que los humanos no pueden, o no tan rápidamente, y es capaz de aprender y cambiar a través de procesos que, en algunos casos, superan el ritmo o el alcance del pensamiento humano? Si los efectos de un arma asistida por IA dependen de la percepción de la IA durante el combate —y de las conclusiones que extraiga de los fenómenos que percibe—, ¿pueden demostrarse los efectos estratégicos de algunas armas solo mediante su uso? Si un competidor entrena a su IA en silencio y en secreto, ¿pueden los líderes saber —fuera de un conflicto— si van por delante o por detrás en una carrera armamentística?

En un conflicto tradicional, la psicología del adversario es un punto crítico al que apunta la acción estratégica. Un algoritmo solo conoce sus instrucciones y objetivos, no su moral ni sus dudas. Debido al potencial de la IA para adaptarse a los fenómenos que encuentra, cuando dos sistemas de armas de IA se despliegan uno contra otro, probablemente ninguna de las partes tenga una comprensión precisa de los resultados que generará su interacción o de sus efectos colaterales. Puede que solo discernan de forma

imprecisa las capacidades del otro y las penalizaciones por entrar en conflicto. Para los ingenieros y constructores, estas limitaciones pueden suponer primas a la velocidad, la amplitud de los efectos y la resistencia, atributos capaces de producir conflictos más intensos y generalizados y, sobre todo, más imprevisibles.

Al mismo tiempo, incluso con la IA, una defensa fuerte es el requisito previo de la seguridad. Su ubicuidad impide el abandono unilateral de la nueva tecnología. Sin embargo, incluso mientras se arman, los gobiernos deberían evaluar e intentar explorar cómo la adición de la lógica de la IA a la experiencia humana de la batalla puede hacer que la guerra sea más humana y precisa, y reflexionar sobre el impacto en la diplomacia y el orden mundial.

La IA y el aprendizaje automático cambiarán las opciones estratégicas y tácticas de los actores al ampliar las capacidades de las clases de armas existentes. La IA no solo permite apuntar a las armas convencionales con mayor precisión, sino también de formas nuevas y poco convencionales, como (al menos en teoría) a un individuo u objeto concreto en vez de a un lugar.¹³ Al analizar grandes cantidades de información, las ciberarmas de IA pueden aprender a penetrar las defensas sin necesidad de que los humanos descubran fallos de software susceptibles de explotarse. Por la misma razón, la IA también puede utilizarse a la defensiva, localizando y reparando fallos antes de que sean explotados. Pero como el atacante puede elegir el objetivo, la IA da a la parte ofensiva una ventaja inherente, e incluso insuperable.

Si un país se enfrenta en combate a un adversario que ha entrenado a su IA para pilotar aviones, tomar decisiones independientes sobre los objetivos y disparar, ¿qué cambios en la táctica, la estrategia o la disposición a recurrir a armas de mayor calibre (o incluso nucleares) producirá la incorporación de esta tecnología?

La IA abre nuevos horizontes de capacidades en el espacio de la información, incluido el ámbito de la desinformación. La IA generativa puede crear grandes cantidades de información falsa pero

verosímil. La desinformación y la guerra psicológica facilitadas por la IA, incluido el uso de personajes, imágenes, vídeos y discursos creados artificialmente, están a punto de generar nuevas e inquietantes vulnerabilidades, especialmente para las sociedades libres. Demostraciones ampliamente difundidas han producido imágenes y vídeos aparentemente realistas de personajes públicos diciendo cosas que nunca han dicho. En teoría, la IA podría utilizarse para determinar las formas más eficaces de transmitir a la gente estos contenidos sintéticos generados por la IA, adaptándolos a sus prejuicios y expectativas. Si la imagen sintética de un dirigente nacional es manipulada por un adversario para fomentar la discordia o emitir directrices engañosas, ¿detectará el público (o incluso otros gobiernos y funcionarios) el engaño a tiempo?

A diferencia de lo que ocurre en el campo de las armas nucleares, estos usos de la IA no van acompañados de una proscripción ampliamente compartida ni de un concepto claro de disuasión (o de grados de escalada). Las potencias de la IA están en condiciones de desplegar máquinas y sistemas con una lógica rápida y un comportamiento emergente y evolutivo para atacar, defender, vigilar, difundir desinformación e identificar y desactivar la IA de los demás.

A medida que las capacidades transformadoras de la IA evolucionan y se extienden, las principales naciones, en ausencia de restricciones verificables, continuarán esforzándose por alcanzar una posición superior.¹⁴ Asumirán que la proliferación de la IA está destinada a ocurrir una vez que se introduzcan nuevas capacidades útiles de IA. En consecuencia, con la ayuda del doble uso civil y militar de esta tecnología y su facilidad de copia y transmisión, los fundamentos y las innovaciones claves de la IA serán, en gran medida, públicos. Cuando las IA están controladas, los controles pueden resultar imperfectos, ya sea porque los avances tecnológicos los dejen obsoletos o porque resulten permeables a un actor determinado. Los nuevos usuarios pueden adaptar los algoritmos subyacentes para objetivos muy diferentes. Una innovación comercial de una sociedad podría ser adaptada por otra con fines de

seguridad o de guerra de la información. Los aspectos estratégicamente más significativos del desarrollo puntero de la IA serán adoptados con frecuencia por los gobiernos para satisfacer sus conceptos de interés nacional.

Los esfuerzos para conceptualizar un equilibrio de poder cibernético y la disuasión de la IA están en su infancia, si acaso. Hasta que no se definan estos conceptos, la planificación tendrá un carácter abstracto. En un conflicto, una parte beligerante puede tratar de abrumar la voluntad de su enemigo mediante el uso, o la amenaza de uso, de un arma cuyos efectos no se conocen bien.

El efecto más revolucionario e imprevisible puede producirse en el punto de encuentro entre la IA y la inteligencia humana. Históricamente, los países que planifican una batalla han sido capaces de comprender, aunque de forma imperfecta, las doctrinas, tácticas y psicología estratégica de sus adversarios. Esto ha permitido desarrollar estrategias y tácticas adversarias, así como un lenguaje simbólico de acciones militares demostrativas, como interceptar un avión que se acerca a una frontera o navegar con un buque por una vía fluvial disputada. Sin embargo, cuando un ejército utiliza la IA para planificar o fijar objetivos —o incluso para ayudar dinámicamente durante una patrulla o un conflicto—, estos conceptos e interacciones familiares pueden volverse extraños porque implicarán la comunicación y la interpretación de una inteligencia que no está familiarizada con sus métodos y tácticas.

Fundamentalmente, el paso a la IA y a las armas y sistemas de defensa asistidos por IA implica una cierta dependencia de —y, en casos extremos, delegación en— una inteligencia de considerable potencial analítico que opera en un paradigma experiencial fundamentalmente diferente. Tal dependencia introducirá riesgos desconocidos o mal comprendidos. Por esta razón, los operadores humanos deben participar y estar en condiciones de supervisar y controlar las acciones de la IA con efectos potencialmente letales. Si este papel humano no evita todos los errores, al menos garantiza la agencia moral y la responsabilidad.

Sin embargo, el reto más profundo puede ser filosófico. Si los aspectos de la estrategia llegan a operar en ámbitos conceptuales y analíticos accesibles a la IA, pero no a la razón humana, se volverán opacos en sus procesos, alcance y significado final. Si los responsables políticos llegan a la conclusión de que la ayuda de la IA en la búsqueda de los patrones más profundos de la realidad es necesaria para comprender las capacidades e intenciones de los adversarios (que pueden disponer de su propia IA) y responder a ellos de manera oportuna, la delegación de decisiones críticas a las máquinas puede llegar a ser inevitable. Es probable que las sociedades arriben a distintos límites instintivos sobre qué delegar y qué riesgos y consecuencias aceptar. Los grandes países no deberían esperar a una crisis para iniciar un diálogo sobre las implicaciones —estratégicas, doctrinales y morales— de estas evoluciones. Si lo hacen, es probable que su impacto sea irreversible. Es imperativo un intento internacional de limitar estos riesgos.

Gestión de la IA

Estas cuestiones deben considerarse y comprenderse antes de enviar sistemas inteligentes a enfrentarse entre sí. Adquieren una urgencia adicional porque el uso estratégico de las capacidades cibernéticas y de IA implica un campo más amplio para las contiendas estratégicas. Se extenderán más allá de los campos de batalla históricos a, en cierto sentido, cualquier lugar conectado a una red digital. Los programas digitales controlan ahora un vasto y creciente reino de sistemas físicos, y un número cada vez mayor de estos sistemas —en algunos casos hasta las cerraduras de las puertas y los frigoríficos— están conectados en red. Esto ha producido un sistema de asombrosa complejidad, alcance y vulnerabilidad.

Para las potencias de IA, es fundamental buscar algún tipo de entendimiento y moderación mutua. En los casos en los que los sistemas y las capacidades se alteran fácilmente y de forma relativamente indetectable mediante un cambio en el código

informático, cada gobierno importante puede asumir que sus adversarios están dispuestos a llevar la investigación, el desarrollo y el despliegue de IA estratégicamente sensible un paso más allá de lo que han reconocido públicamente o incluso prometido en privado. Desde un punto de vista puramente técnico, las líneas que separan el reconocimiento, la selección de objetivos y la acción autónoma letal de la IA se cruzan con relativa facilidad, lo cual hace que la búsqueda de sistemas de restricción y verificación mutuas sea tan difícil como imperativa.

La búsqueda de seguridad y moderación deberá enfrentarse a la naturaleza dinámica de la IA. Una vez liberadas en el mundo, las armas cibernéticas facilitadas por la IA pueden ser capaces de adaptarse y aprender mucho, más allá de sus objetivos previstos; las propias capacidades del arma podrían cambiar a medida que la IA reacciona a su entorno. Si las armas son capaces de cambiar de forma diferente en alcance o tipo a lo que sus creadores anticiparon o amenazaron, los cálculos de disuasión y escalada pueden volverse ilusorios. Porque la gama de actividades que una IA es capaz de emprender, tanto en la fase inicial de diseño como durante la fase de despliegue, puede tener que ajustarse para que un humano conserve la capacidad de supervisar y apagar o redirigir un sistema que ha empezado a desviarse. Para evitar resultados inesperados y potencialmente catastróficos, estas restricciones deben ser recíprocas.

Será difícil definir las limitaciones de las capacidades cibernéticas y de IA, y será difícil detener la proliferación. Las capacidades desarrolladas y utilizadas por las grandes potencias pueden caer en manos de terroristas y delincuentes. Del mismo modo, las naciones más pequeñas, no poseedoras de armas nucleares y con una capacidad limitada de armamento convencional, tienen la capacidad de ejercer una gran influencia invirtiendo en arsenales cibernéticos y de IA de vanguardia.

Inevitablemente, los países delegarán tareas discretas y no letales en algoritmos de IA (algunos operados por entidades privadas),

incluida la realización de funciones defensivas que detecten y eviten intrusiones en el ciberespacio. La «superficie de ataque» de una sociedad digital altamente interconectada será demasiado amplia para que los operadores humanos puedan defenderla manualmente. A medida que muchos aspectos de la vida humana se trasladan a Internet y las economías siguen digitalizándose, una ciber IA malintencionada podría perturbar sectores enteros. Los países, las empresas e incluso los individuos, deberían invertir en sistemas de seguridad para protegerse de estas situaciones.

La forma más extrema de esta protección consistirá en cortar las conexiones de red y desconectar los sistemas. Para las naciones, la desconexión podría convertirse en la forma definitiva de defensa. A falta de medidas tan extremas, solo la IA será capaz de llevar a cabo ciertas funciones vitales de ciberdefensa, en parte debido a la vasta extensión del ciberespacio y a la casi infinita gama de acciones posibles dentro de él. Por lo tanto, las capacidades defensivas más significativas en este ámbito estarán probablemente fuera del alcance de todas las naciones, excepto de unas pocas.

Más allá de los sistemas de defensa habilitados por IA, se encuentra la categoría de capacidades más controvertida: los sistemas de armas autónomas letales, generalmente entendidos como sistemas que, una vez activados, pueden seleccionar y atacar objetivos sin intervención humana.¹⁵ La cuestión clave en este ámbito es la supervisión humana y la capacidad de intervención humana oportuna.

Un sistema autónomo puede tener a un ser humano «*on the loop*»¹⁶, supervisando sus actividades pasivamente, o «*in the loop*», con la necesaria autorización humana para determinadas acciones. A menos que se restrinja mediante un acuerdo mutuo observable y verificable, estos sistemas de armas pueden llegar a abarcar estrategias y objetivos enteros —como defender una frontera o lograr un resultado concreto contra un adversario— y operar sin una participación humana sustancial. En estos ámbitos, es imperativo garantizar un papel adecuado para el juicio humano en la

supervisión y dirección del uso de la fuerza. Estas limitaciones solo tendrán un significado limitado si son adoptadas unilateralmente por una nación o un pequeño grupo de naciones. Los gobiernos de los países tecnológicamente avanzados deberían explorar los retos de la restricción mutua apoyada por una verificación ejecutable.¹⁷

La IA aumenta el riesgo inherente de anticipación y de que su uso prematuro desemboque en un conflicto. Un país que teme que su adversario esté desarrollando capacidades automáticas puede tratar de adelantarse a él: si el ataque «tiene éxito», quizás no haya forma de saber si estaba justificado. Para evitar una escalada imprevista, las grandes potencias deberían mantener su competición dentro de un marco de límites verificables. La negociación no solo debe centrarse en moderar una carrera armamentística, sino también en asegurarse de que ambas partes saben, en términos generales, lo que está haciendo la otra. Pero ambas partes deben esperar (y planificar en consecuencia) que la otra retenga sus secretos más sensibles desde el punto de vista de la seguridad. La confianza nunca será total. Pero como demostraron las negociaciones sobre armamento nuclear durante la Guerra Fría, eso no significa que no pueda alcanzarse un cierto grado de entendimiento.

Planteamos estas cuestiones en un esfuerzo por definir los retos que la IA introduce en la estrategia. A pesar de todas sus ventajas, los tratados (y los mecanismos de comunicación, aplicación y verificación que los acompañan) que llegaron a definir la era nuclear no eran inevitables desde el punto de vista histórico. Fueron producto de la acción humana y del reconocimiento mutuo del peligro y la responsabilidad.

Impacto en las tecnologías civiles y militares

Tradicionalmente, tres cualidades han facilitado la separación de los ámbitos militar y civil: la diferenciación tecnológica, el control concentrado y la magnitud del efecto. Las tecnologías con aplicaciones exclusivamente militares o exclusivamente civiles se

describen como diferenciadas. El control concentrado se refiere a las tecnologías que un gobierno puede gestionar fácilmente, por oposición a las que se extienden con facilidad y escapan así al control gubernamental. Por último, la magnitud del efecto alude al potencial destructivo de una tecnología.

A lo largo de la historia, muchas tecnologías han sido de doble uso. Otras se han extendido fácil y ampliamente, y algunas han tenido un enorme potencial destructivo. Sin embargo, hasta ahora ninguna ha tenido las tres características: doble uso, fácil difusión y potencial destructivo sustancial. Los ferrocarriles que llevaban mercancías al mercado eran los mismos que llevaban soldados a la batalla, pero no tenían potencial destructivo. Las tecnologías nucleares suelen ser de doble uso y pueden generar una enorme capacidad destructiva, pero su complicada infraestructura permite un control gubernamental relativamente seguro. Un rifle de caza puede ser de uso generalizado y tener aplicaciones militares y civiles, pero su capacidad limitada impide a su portador infligir destrucción a nivel estratégico.

La IA rompe este paradigma. Es de doble uso. Se propaga con facilidad al no ser, en esencia, más que líneas de código: la mayoría de los algoritmos (con algunas excepciones dignas de mención) pueden ejecutarse en un solo ordenador o en pequeñas redes, lo cual significa que los gobiernos tienen dificultades para controlar la tecnología mediante el control de la infraestructura. Por último, las aplicaciones de la IA tienen un importante potencial destructivo. Esta constelación relativamente única de cualidades, unida al amplio abanico de partes interesadas, genera retos estratégicos de una complejidad novedosa.

Las armas habilitadas por IA (*AI-enabled weapons*)¹⁸ pueden permitir a los adversarios lanzar ataques digitales a una velocidad excepcional, y aceleran drásticamente la capacidad humana para explotar las vulnerabilidades digitales. De este modo, es posible que un Estado no tenga tiempo para evaluar las señales de un ataque entrante. Si un Estado tiene los medios, puede optar por responder

casi simultáneamente, antes de que el ataque pueda producirse por completo, construyendo un sistema habilitado por IA que busque ataques y lo habilite para contraatacar.¹⁹ Para el bando contrario, la noticia de la existencia de un sistema de este tipo y el conocimiento de que podría actuar sin previo aviso puede servir de acicate para una construcción y planificación adicionales, que pueden incluir el desarrollo de una tecnología paralela o basada en algoritmos diferentes. A menos que se tenga cuidado de desarrollar un concepto común de los límites, la compulsión de actuar primero puede anular la necesidad de actuar con prudencia —como ocurría a principios del siglo XX—, si es que los seres humanos participan realmente en este tipo de decisiones.

En el mercado de valores, las sofisticadas empresas denominadas *quant* han reconocido que los algoritmos de IA pueden detectar patrones de mercado y reaccionar con una rapidez que supera incluso la del operador más hábil. En consecuencia, estas empresas han delegado en estos algoritmos el control de determinados aspectos de sus operaciones con valores. En muchos casos, estos sistemas algorítmicos pueden superar los beneficios humanos por un margen sustancial. Sin embargo, en ocasiones cometen errores de cálculo graves, potencialmente muy superiores al peor error humano.

En el mundo financiero, estos errores devastan las carteras pero no se cobran vidas. En el ámbito estratégico, sin embargo, un fallo algorítmico análogo a un «*flash crash*» podría ser catastrófico. Si la defensa estratégica en el ámbito digital requiere una ofensiva táctica, si una de las partes se equivoca en sus cálculos o en sus acciones, podría desencadenarse inadvertidamente un patrón de escalada.

Los intentos de incorporar estas nuevas capacidades a un concepto definido de estrategia y equilibrio internacional se complican por el hecho de que los conocimientos necesarios para la preeminencia tecnológica ya no se concentran exclusivamente en el gobierno. Un amplio abanico de actores e instituciones participan en

la creación de tecnología con implicaciones estratégicas, desde los tradicionales contratistas gubernamentales hasta los inventores individuales, los empresarios, las *start-ups* y los laboratorios de investigación privados. No todos considerarán que sus misiones son intrínsecamente compatibles con los objetivos nacionales definidos por el gobierno federal. Un proceso de educación mutua entre la industria, el mundo académico y el gobierno puede ayudar a salvar esta distancia y garantizar que los principios claves de las implicaciones estratégicas de la IA se comprendan en un marco conceptual común. Pocas épocas se han enfrentado a un reto estratégico y tecnológico tan complejo y con tan poco consenso sobre la naturaleza del desafío o incluso sobre el vocabulario necesario para debatirlo.

El reto sin resolver de la era nuclear fue que la humanidad desarrolló una tecnología para la que los estrategas no pudieron encontrar una doctrina operativa viable. El dilema de la era de la IA será diferente: su tecnología definitoria será ampliamente adquirida, dominada y empleada. La consecución de una moderación estratégica mutua —o incluso el logro de una definición común de moderación— será más difícil que nunca, tanto desde el punto de vista conceptual como práctico.

La gestión de las armas nucleares, el empeño de medio siglo, sigue siendo incompleta y fragmentaria. Sin embargo, el reto de evaluar el equilibrio nuclear era relativamente sencillo. Se podían contar las cabezas nucleares y se conocía su rendimiento. Por el contrario, las capacidades de las IA no son fijas; son dinámicas. A diferencia de las armas nucleares, las IA son difíciles de rastrear: una vez entrenadas, pueden copiarse fácilmente y funcionar en máquinas relativamente pequeñas. Y detectar su presencia o verificar su ausencia es difícil o imposible con la tecnología actual. En esta era, la disuasión surgirá probablemente de la complejidad: de la multiplicidad de vectores a través de los cuales puede viajar un ataque posibilitado por la IA y de la velocidad de las posibles respuestas de la IA.

Para gestionar la IA, los estrategas deben considerar cómo podría integrarse en un modelo responsable de relaciones internacionales. Antes de desplegar las armas, los estrategas deben comprender el efecto iterativo de su uso, el potencial de escalada y las vías de desescalada. Es esencial una estrategia de uso responsable, que incluya principios de contención. Los responsables políticos deben esforzarse por abordar simultáneamente el armamento, las tecnologías y estrategias defensivas, junto con el control de armamentos, en lugar de considerarlos pasos cronológicamente distintos y funcionalmente antagónicos. Deben formularse las doctrinas y tomarse las decisiones antes de la utilización de este armamento.

¿Cuáles serán entonces los requisitos de moderación? La tradicional imposición de moderación a la capacidad es un punto de partida obvio. Durante la Guerra Fría, este planteamiento supuso un cierto progreso, al menos simbólico. Algunas capacidades se restringieron (las cabezas nucleares, por ejemplo); otras (como las categorías de misiles de alcance intermedio) se prohibieron rotundamente. Pero ni restringir las capacidades subyacentes de las IA ni limitar su número sería totalmente compatible con el uso civil generalizado de la tecnología y su continua evolución. Habrá que estudiar restricciones adicionales, centradas en las capacidades de aprendizaje y de selección de objetivos de las IA.

En una decisión que ha previsto en parte este desafío, Estados Unidos ha distinguido entre *AI-enabled weapons* (armas habilitadas por IA), que hacen que la guerra dirigida por humanos sea más precisa, más letal y más eficaz, y *AI weapons* (armas con IA), que toman decisiones letales de forma autónoma respecto a los operadores humanos. Estados Unidos ha declarado su objetivo de restringir el uso a la primera categoría. Aspira a un mundo en el que nadie, ni siquiera los propios Estados Unidos, posea la segunda.²⁰ Esta distinción es acertada. Al mismo tiempo, la capacidad de la tecnología para aprender y, por tanto, evolucionar, podría hacer que las restricciones sobre determinadas capacidades resultaran

insuficientes. Será fundamental definir la naturaleza y la forma de restringir las armas dotadas de IA, y garantizar que la restricción sea mutua.

En los siglos XIX y XX, las naciones impusieron restricciones a ciertas formas de guerra: el uso de armas químicas, por ejemplo, y los ataques desproporcionados contra civiles. A medida que las armas de inteligencia artificial hacen posibles nuevas categorías de actividades o potencian las antiguas, las naciones del mundo deben tomar decisiones urgentes sobre lo que es compatible con los conceptos de dignidad humana inherente y agencia moral. La seguridad exige anticiparse a lo que está por venir, no limitarse a reaccionar ante lo ya existente.

He aquí el dilema planteado por la tecnología armamentística relacionada con la IA: mantener la investigación y el desarrollo es esencial para la supervivencia nacional; sin ello, perderemos competitividad y relevancia comercial. Pero la proliferación inherente a la nueva tecnología ha frustrado hasta ahora cualquier intento de restricción negociada, incluso conceptualmente.

Una vieja búsqueda en un mundo nuevo

Todos los países tecnológicamente avanzados deben comprender que se encuentran en el umbral de una transformación estratégica tan importante como la aparición de las armas nucleares, pero con efectos más diversos, difusos e impredecibles. Cada sociedad que esté avanzando en las fronteras de la IA debería intentar convocar un organismo a nivel nacional para considerar los aspectos de defensa y seguridad de la IA y aunar las perspectivas de los diversos sectores que darán forma a la creación y despliegue de esta. A este organismo deberían confiársele dos funciones: garantizar la competitividad con el resto del mundo y, al mismo tiempo, coordinar la investigación sobre cómo prevenir o al menos limitar una escalada o crisis no deseada. Sobre esta base, será esencial algún tipo de negociación con aliados y adversarios.

Si se explora esta dirección, resultará esencial que las principales potencias mundiales de la IA —Estados Unidos y China— acepten esta realidad. Podrían llegar a la conclusión de que, sean cuales sean las demás contiendas que pueda deparar un periodo de rivalidad emergente, Estados Unidos y China deberían buscar el consenso para no entrar en una guerra tecnológicamente avanzada entre ellos. Podría confiarse a una unidad o a un subconjunto de funcionarios de alto rango de cada gobierno la tarea de vigilar e informar directamente a su presidente sobre los peligros incipientes y la forma de evitarlos. En el momento de escribir estas líneas, este no es un empeño coincidente con el sentimiento público de ninguna de las dos naciones. Sin embargo, cuanto más tiempo pasen las dos potencias tratándose como rivales institucionalizados sin entablar un diálogo de este tipo, mayores serán las posibilidades de que se produzca un accidente en el que ambas partes se vean empujadas por sus tecnologías y calendarios de despliegue a una crisis que ninguna de las dos busca y ambas llegan a lamentar, y que puede incluir un conflicto militar a escala mundial.

La paradoja de un sistema internacional es que cada potencia se ve impulsada a actuar —de hecho, debe actuar— para maximizar su propia seguridad. Sin embargo, para evitar una serie constante de crisis, cada una debe aceptar cierto sentido de la responsabilidad en el mantenimiento de la paz general. Y este proceso implica un reconocimiento de los límites. El planificador militar o el responsable de seguridad pensará (no incorrectamente) en términos de los peores escenarios y priorizará la adquisición de capacidades para hacerles frente. El estadista (que puede ser uno y el mismo) está obligado a considerar cómo se utilizarán estas capacidades y qué aspecto tendrá el mundo después.

En la era de la IA, habrá que adaptar la lógica estratégica de siempre. Tendremos que superar, o al menos moderar, el impulso hacia el automatismo antes de que se produzca una catástrofe. Debemos evitar que las IA que operan más rápido que los decisores humanos emprendan acciones irreversibles con consecuencias

estratégicas. Habrá que automatizar las defensas sin ceder los elementos esenciales del control humano. La ambigüedad inherente al dominio —combinada con las cualidades dinámicas y emergentes de la IA y su facilidad de difusión— complicará las evaluaciones. En épocas anteriores, solo un puñado de grandes potencias o superpotencias tenían la responsabilidad de frenar sus capacidades destructivas y evitar catástrofes. Pronto, la proliferación puede llevar a que muchos más actores asuman una tarea similar.

Los líderes de esta era pueden aspirar a seis tareas principales en el control de sus arsenales, con su amplia y dinámica combinación de capacidades convencionales, nucleares, cibernéticas y de IA.

En primer lugar, los dirigentes de las naciones rivales y adversarias deben estar dispuestos a hablar entre sí con regularidad, como hicieron sus predecesores durante la guerra fría, sobre las formas de guerra que no desean librar. Para contribuir a este esfuerzo, Washington y sus aliados deben organizarse en torno a intereses y valores que identifiquen como comunes, inherentes e inviolables y que engloben las experiencias de las generaciones que alcanzaron la mayoría de edad al final de la Guerra Fría o después de ella.

En segundo lugar, los enigmas sin resolver de la estrategia nuclear deben recibir una nueva atención y ser reconocidos como lo que son: uno de los grandes retos humanos estratégicos, técnicos y morales. Durante muchas décadas, los recuerdos de unas Hiroshima y Nagasaki humeantes obligaron a reconocer los asuntos nucleares como una empresa única y grave. Como dijo el ex secretario de Estado George Shultz al Congreso en 2018: «Me temo que la gente ha perdido esa sensación de pavor». Los líderes de los países con armas nucleares deben reconocer su responsabilidad de trabajar juntos para evitar la catástrofe.

En tercer lugar, las principales potencias cibernéticas y de IA deberían esforzarse por definir sus doctrinas y límites (aunque no todos sus aspectos se anuncien públicamente) e identificar los puntos de correspondencia entre sus doctrinas y las de las potencias

rivales. Para que la disuasión predomine sobre el uso, la paz sobre el conflicto y el conflicto limitado sobre el conflicto general, estos asuntos deberán entenderse y definirse en términos que reflejen los aspectos distintivos de la cibernética y la IA.

En cuarto lugar, los Estados poseedores de armas nucleares deberían comprometerse a llevar a cabo sus propias revisiones internas de sus sistemas de mando y control y de alerta temprana. Estas revisiones a prueba de fallos identificarían los pasos para reforzar las protecciones contra las amenazas cibernéticas y el uso no autorizado, inadvertido o accidental de armas de destrucción masiva. Estas revisiones también deberían incluir opciones para prevenir ciberataques contra los activos nucleares de mando y control o de alerta temprana.

En quinto lugar, los países —especialmente los tecnológicamente grandes— deberían crear métodos sólidos y aceptados para maximizar el tiempo de decisión durante los periodos de mayor tensión y en situaciones extremas. Este debe ser un objetivo conceptual común, especialmente entre adversarios, que conecte los pasos inmediatos y a largo plazo para gestionar la inestabilidad y construir la seguridad mutua. En una crisis, los seres humanos deben asumir la responsabilidad final sobre el despliegue o no de armas avanzadas. Especialmente los adversarios deberían esforzarse por acordar un mecanismo capaz de garantizar que las decisiones que puedan resultar irrevocables se tomen a un ritmo que favorezca el pensamiento y la deliberación humanos, y la supervivencia.^{[21](#)}

Por último, las principales potencias de IA deberían plantearse cómo limitar la proliferación continuada de IA militar o si emprender un esfuerzo sistémico de no proliferación, respaldado por la diplomacia y la amenaza de la fuerza. ¿Quiénes son los aspirantes a adquirir la tecnología que la utilizarían con fines destructivos inaceptables? ¿Qué armas específicas de IA justifican esta preocupación? ¿Y quién hará cumplir la línea roja? Las potencias nucleares establecidas exploraron este concepto para la proliferación nuclear, con un éxito desigual. Si se permite que una nueva

tecnología, perturbadora y potencialmente destructiva, transforme los ejércitos de los gobiernos más inveteradamente hostiles o moralmente incontrolados del mundo, el equilibrio estratégico puede resultar difícil de alcanzar y el conflicto, entonces, incontrolable.

Debido al carácter de doble uso de la mayoría de las tecnologías de IA, tenemos el deber para con nuestra sociedad de mantenernos a la vanguardia de la investigación y el desarrollo. Pero esto nos obligará igualmente a comprender los límites. Si llega una crisis, será demasiado tarde para empezar a debatir estas cuestiones. Una vez empleada en un conflicto militar, la velocidad de la tecnología prácticamente garantiza que impondrá resultados a un ritmo más rápido que el que puede desplegar la diplomacia. Las grandes potencias deben entablar un debate sobre las armas cibernéticas y con inteligencia artificial, aunque solo sea para desarrollar un vocabulario común de conceptos estratégicos y un cierto sentido de las líneas rojas de cada una. La voluntad de lograr una contención mutua de las capacidades más destructivas no debe esperar a que surja la tragedia. Cuando la humanidad se lance a competir en la creación de armas nuevas, evolutivas e inteligentes, la historia no perdonará que se renuncie a establecer límites. En la era de la inteligencia artificial, la búsqueda permanente de la ventaja nacional debe basarse en una ética de preservación humana.

¹ John P. Glennon, ed., *Foreign Relations of the United States*, vol. 19, National Security Policy, 1955-1957 (Washington, DC: US Government Printing Office, 1990), 61.

² Véase Henry A. Kissinger, *Armas nucleares y política internacional*, traducción de Rafael Cremades. Ediciones Rialp, Madrid, 1962.

³ Véase, por ejemplo, Department of Defense, «America's Nuclear Triad», <https://www.defense.gov/Experience/Americas-Nuclear-Triad/>.

⁴ Véase, por ejemplo, Defense Intelligence Agency, «Russia Military Power: Building a Military to Support Great Power Aspirations» (unclassified), 2017, 26-27, <https://www.dia.mil/Portals/27/Documents/News/Military%20Power%20Publicatio>

[ns/Russia%20Military%20Power%20Report%202017.pdf](#); Anthony M. Barrett, «False Alarms, True Dangers? Current and Future Risks of Inadvertent U.S.-Russian Nuclear War», 2016, https://www.rand.org/content/dam/rand/pubs/perspectives/PE100/PE191/RAND_PE191.pdf; David E. Hoffman, *The Dead Hand: The Untold Story of the Cold War Arms Race and Its Dangerous Legacy* (New York: Doubleday, 2009).

⁵ Por ejemplo, el malware NotPetya, desplegado por operadores rusos contra instituciones financieras y agencias gubernamentales ucranianas en 2017, acabó extendiéndose mucho más allá de las entidades objetivo en Ucrania, a centrales eléctricas, hospitales, proveedores de transporte y logística y empresas energéticas de otros países, incluida la propia Rusia. Como dijeron el senador Angus King y el representante Mike Gallagher, presidentes de la Comisión Cyberspace Solarium, de Estados Unidos, en su informe de marzo de 2020: «Como una infección en el torrente sanguíneo, el malware se extendió a lo largo de las cadenas de suministro globales.» Véase la página 8 del Informe de la Comisión Cyberspace Solarium de Estados Unidos, https://drive.google.com/file/d/1ryMCIL_dZ30QyJFqFkkf10MxIXJGT4yv/view.

⁶ Véase Andy Greenberg, *Sandworm: A New Era of Cyberwar and the Hunt for the Kremlin's Most Dangerous Hackers* (New York: Doubleday, 2019); Fred Kaplan, *Dark Territory: The Secret History of Cyber War* (New York: Simon & Schuster, 2016).

⁷ Véase Richard Clarke y Robert K. Knake, *The Fifth Domain: Defending Our Country, Our Companies, and Ourselves in the Age of Cyber Threats* (New York: Penguin Press, 2019).

⁸ Véase, por ejemplo, Summary: Department of Defense Cyber Strategy 2018, https://media.defense.gov/2018/Sep/18/2002041658/-1/-1/1/CYBER_STRATEGY_SUMMARY_FINAL.PDF.

⁹ Para una visión general ilustrativa, véase Eric Schmidt, Robert Work, et al., *Final Report: National Security Commission on Artificial Intelligence*, March 2021, <https://www.nscai.gov/2021-final-report>; Christian Brose, *The Kill Chain: Defending America in the Future of High-Tech Warfare* (New York: Hachette Books, 2020); Paul Scharre, *Army of None: Autonomous Weapons and the Future of War* (New York: W. W. Norton, 2018).

¹⁰ Will Roper, «AI Just Controlled a Military Plane for the First Time Ever», *Popular Mechanics*, December 16, 2020, <https://www.popularmechanics.com/military/aviation/a34978872/artificial-intelligence-controls-u2-spy-plane-air-force-exclusive>.

¹¹ Véase, por ejemplo, «Automatic Target Recognition of Personnel and Vehicles from an Unmanned Aerial System Using Learning Algorithms,» SBIR/STTR (Small Business Innovation Research and Small Business Technology Transfer programs), November 29, 2017 («Objetivo: Desarrollar un sistema que pueda integrarse y desplegarse en un sistema aéreo no tripulado de clase 1 o clase 2... para detectar, reconocer, clasificar, identificar... y apuntar automáticamente a personal y plataformas terrestres u otros objetivos de interés»), <https://www.sbir.gov/sbirsearch/detail/1413823>; Gordon Cooke, «Magic Bullets: The Future of Artificial Intelligence in Weapons Systems», Army AL&T, junio de 2019, https://www.army.mil/article/223026/magic_bullets_the_future_of_artificial_intelligence_in_weapons_systems.

¹² Scharre, *Army of None*, 102-119.

¹³ Véase, por ejemplo, United States White House Office, «National Strategy for Critical and Emerging Technologies», October, 2020, <https://www.hsdl.org/?view&did=845571>; Central Committee of the Communist Party of China, *14th Five-Year Plan for Economic and Social Development and 2035 Vision Goals*, March 2021; Xi Jinping, «Strive to Become the World's Major Scientific Center and Innovation Highland», discurso pronunciado en la Conferencia de Académicos de la Academia China de Ciencias y la Academia China de Ingeniería, 28 de mayo de 2018, en Qishui, marzo de 2021; Comisión Europea, Libro Blanco sobre Inteligencia Artificial: Un enfoque europeo para la excelencia y la confianza, marzo de 2020.

¹⁴ Véase, por ejemplo, Department of Defense Directive 3000.09, «Autonomía en los sistemas de armas», rev. May 8, 2017, <https://www.esd.whs.mil/portals/54/documents/dd/issuances/dodd/300009p.pdf>.

¹⁵ Véase, por ejemplo, Schmidt, Work, et al., Final Report, 10, 91-101; Department of Defense, «DOD Adopts Ethical Principles for Artificial Intelligence», February 24, 2020, <https://www.defense.gov/Newsroom/Releases/Release/Article/2091996/dod-adopts-ethical-principles-for-artificial-intelligence>; Defense Innovation Board, «AI Principles: Recommendations on the Ethical Use of Artificial Intelligence by the Department of Defense», [https://admin.govexec.com/media/dib_ai_principles_-_supporting_document_-_embargoed_copy_\(oct_2019\).pdf](https://admin.govexec.com/media/dib_ai_principles_-_supporting_document_-_embargoed_copy_(oct_2019).pdf).

¹⁶ Véase, por ejemplo, Schmidt, Work, et al., Final Report, 9, 278-282.

¹⁷ *N de la T.*: Se ha optado por su formulación original, pues son los términos en inglés habitualmente utilizados en IA en todo el mundo. En español sería algo así

como «sobre o encima del bucle» (on the loop) para los sistemas capaces de discriminar objetivos, fijarlos e incluso atacarlos por sí mismos, aunque bajo supervisión humana pasiva (que solo puede abortar la acción) y «en el o dentro del bucle» o «humano en el circuito» (in the loop) para las armas que requieren de la participación humana activa, tanto para seleccionar los objetivos como para atacarlos.

¹⁸ *N. de la T.*: En la literatura especializada en español, «AI-enabled weapons» se traduce indistintamente como «armas dotadas de IA» o «armas habilitadas por IA». En todo caso, se refiere a armas que incorporan la inteligencia artificial, pero son controladas por humanos, a diferencia de las «AI weapons» («armas con IA»), completa y letalmente autónomas.

¹⁹ Scharre, *Army of None*, 226-228.

²⁰ Véase, por ejemplo, Congressional Research Service, «Defense Primer: U.S. Policy on Lethal Autonomous Weapon Systems», actualizado el 1 de diciembre de 2020, <https://crsreports.congress.gov/product/pdf/IF/IF11150>; Department of Defense Directive 3000.09, § 4(a); Schmidt, Work, et al., Final Report, 92-93.

²¹ Versiones de estos conceptos se exploraron inicialmente en William J. Perry, Henry A. Kissinger y Sam Nunn, «Building on George Shultz's Vision of a World Without Nukes», Wall Street Journal, 23 de mayo de 2021, <https://www.wsj.com/articles/building-on-george-shultzs-vision-of-a-world-without-nukes-11616537900>.

IA e identidad humana

En una era en la cual las máquinas realizan cada vez más tareas de las que antes solo eran capaces los humanos, ¿qué constituirá entonces nuestra identidad como seres humanos? Como se ha explicado en capítulos anteriores, la IA ampliará lo que conocemos de la realidad. Modificará nuestra forma de comunicarnos, trabajar en red y compartir información. Transformará las doctrinas y estrategias que desarrollamos e implementamos. Cuando dejemos de explorar y modelar la realidad por nuestra cuenta —cuando recurramos a la IA como complemento de nuestras percepciones y pensamientos—, ¿cómo llegaremos a vernos a nosotros mismos y a nuestro papel en el mundo? ¿Cómo conciliaremos la IA con conceptos como la autonomía y la dignidad humanas?

En épocas anteriores, los seres humanos se han situado en el centro de la historia. Aunque la mayoría de las sociedades reconocen la imperfección humana, han sostenido que las capacidades y experiencias humanas constituyen la culminación de lo que los seres mortales pueden aspirar a lograr en el mundo. De hecho, han celebrado a individuos que han ejemplificado los pináculos del espíritu humano, ilustrando cómo deseamos vernos a nosotros mismos. Estos héroes han variado de una sociedad a otra y de una época a otra —líderes, exploradores, inventores, mártires—, pero todos han encarnado aspectos de los logros humanos y, por tanto, de la singularidad humana. En la era moderna, nuestra veneración por los héroes se ha centrado en los pioneros del ejercicio de la razón —astronautas, inventores, empresarios, líderes políticos— que exploran y organizan nuestra realidad.

Ahora estamos entrando en una era en la que a la IA —una creación humana— se le confían cada vez más tareas que antes habrían sido realizadas, o intentadas, por mentes humanas. A medida que la IA ejecuta estas tareas, y produce resultados que se aproximan y a veces superan a los de la inteligencia humana, desafía un atributo definitorio de lo que significa ser humano. Además, la IA es capaz de aprender, evolucionar y ser «mejor» (según la función objetivo que se le haya asignado). Este aprendizaje dinámico permite a la IA alcanzar resultados complejos que, hasta ahora, estaban reservados a los seres humanos y a las organizaciones humanas.

Con el auge de la IA, cambiarán las definiciones del papel humano, las aspiraciones humanas y la realización humana. ¿Qué cualidades humanas celebrará esta era? ¿Cuáles serán sus principios rectores? A las dos formas tradicionales de conocer el mundo, la fe y la razón, la IA añade una tercera. Este cambio pondrá a prueba —y, en algunos casos, transformará— nuestros supuestos básicos sobre el mundo y nuestro lugar en él. La razón no solo revolucionó las ciencias, sino que también alteró nuestra vida social, nuestras artes y nuestra fe. Bajo su escrutinio, cayó la jerarquía del feudalismo y surgió la democracia, la idea de que las personas que razonan deben dirigir su propio gobierno. Ahora, la IA volverá a poner a prueba los principios en los que se basa nuestra autocomprensión.

En una era en la que la realidad puede ser predicha, aproximada y simulada por una IA capaz de evaluar lo que es relevante para nuestras vidas, predecir lo que vendrá después y decidir qué hacer, el papel de la razón humana cambiará. Con ello, cambiará también el sentido de nuestros fines individuales y sociales. En algunos ámbitos, la IA puede aumentar la razón humana. En otros, la IA puede provocar en los humanos la sensación de ser tangenciales al proceso primario que gobierna una situación. Para el conductor cuyo vehículo selecciona un carril o una ruta diferentes basándose en un cálculo inexplicable —de hecho, tácito—, para la persona a quien se le concede o deniega un crédito a partir de una revisión facilitada

por la IA, para el solicitante de empleo a quien se le pide o no una entrevista sobre la base de un proceso similar, y para el académico al que un modelo de IA le dice la respuesta más probable antes de que su investigación haya comenzado en serio, la experiencia puede resultar eficiente pero no siempre satisfactoria. Para los humanos acostumbrados a la agencia, la centralidad y el monopolio de la inteligencia compleja, la IA supondrá un reto para la autopercepción.

Los avances que hemos considerado hasta ahora son ejemplos de los muchos modos en que la IA está cambiando nuestra forma de interactuar con el mundo y, por tanto, de concebirnos a nosotros mismos y nuestro papel en él. La IA hace predicciones, tales como la probabilidad de que una persona tenga cáncer de mama en fase inicial; toma decisiones, como qué jugada hacer en ajedrez; destaca y filtra información, como qué películas ver o qué inversiones mantener; y genera textos similares a los humanos, desde frases a párrafos y documentos enteros. A medida que aumenta la sofisticación de estas capacidades, la IA se convierte rápidamente en lo que la mayoría de la gente considera creativo o experto. El hecho de que la IA sea capaz de hacer ciertas predicciones o tomar ciertas decisiones, o de generar cierto material, no indica por sí mismo una sofisticación similar a la de los humanos. Pero en muchos casos, los resultados son comparables o superiores a los que antes solo producían los humanos.

Consideremos el texto que modelos generativos como GPT-3 son capaces de crear. Casi cualquier persona con estudios primarios puede hacer un trabajo razonable prediciendo las posibles terminaciones de una frase. Pero escribir documentos y código, algo que GPT-3 puede hacer, requiere sofisticadas habilidades que los humanos pasan años desarrollando en la educación superior. Así pues, los modelos generativos empiezan a cuestionar nuestra creencia de que tareas como completar frases son distintas y más sencillas que escribir. A medida que mejoren los modelos generativos, la IA dará lugar a nuevas percepciones tanto de la

singularidad como del valor relativo de las capacidades humanas. ¿Dónde nos dejará esto?

Con percepciones de la realidad complementarias a las humanas, la IA puede surgir como un socio eficaz para las personas. En el descubrimiento científico, el trabajo creativo, el desarrollo de software y otros campos comparables, puede ser muy beneficioso contar con un interlocutor con una percepción diferente. Pero esta colaboración exigirá que los humanos se adapten a un mundo en el que nuestra razón no es la única forma —y quizá tampoco la más informativa— de conocer o navegar por la realidad. Esto presagia un cambio en la experiencia humana más significativo que cualquier otro que se haya producido en casi seis siglos, desde la aparición de la imprenta de tipos móviles.

Las sociedades tienen dos opciones: reaccionar y adaptarse poco a poco, o iniciar intencionadamente un diálogo, recurriendo a todos los elementos de la empresa humana, dirigido a definir el papel de la IA y, al hacerlo, definir el nuestro. El primer camino lo encontraremos por defecto. El segundo requerirá un compromiso consciente entre líderes y filósofos, científicos y humanistas y otros grupos.

En última instancia, los individuos y las sociedades deberán decidir qué aspectos de la vida reservan a la inteligencia humana y cuáles ceden a la IA o a la colaboración entre humanos e IA. La colaboración entre humanos e IA no se da entre iguales. En última instancia, los humanos construyen y dirigen la IA. Pero a medida que nos habituemos a la IA y dependamos de ella, restringirla puede resultar más costoso y psicológicamente difícil o incluso más complicado desde el punto de vista técnico. Nuestra tarea consistirá en comprender las transformaciones que la IA aporta a la experiencia humana, los retos que plantea a la identidad humana y qué aspectos de estos avances requieren regulación o contrapeso por parte de otros compromisos humanos. Trazar un futuro humano pasa por definir el papel del ser humano en la era de la IA.

Transformar la experiencia humana

Para algunos, la experiencia de la IA será fortalecedora. En la mayoría de las sociedades, una pequeña pero creciente cohorte entiende de IA. Para estos individuos —las personas que la construyen, la entrenan, le asignan tareas y la regulan— y para los responsables políticos y empresarios con asesores técnicos a su disposición, la asociación debería ser gratificante, aunque a veces sorprendente. De hecho, en muchos campos, la experiencia de superar la razón tradicional mediante tecnología especializada, como en los casos de los avances de la IA en medicina, biología, química y física, resultará a menudo satisfactoria.

Quienes carecen de conocimientos técnicos, o participan en procesos gestionados por la IA principalmente como consumidores, también encontrarán con frecuencia estos procesos gratificantes, como en el caso de una persona ocupada que puede leer o consultar su correo electrónico mientras viaja en un coche autónomo. De hecho, la integración de la IA en los productos de consumo distribuirá ampliamente los beneficios de la tecnología. Sin embargo, la IA también operará redes y sistemas no diseñados para el beneficio de ningún usuario individual específico y que están más allá del control de cualquier usuario individual. En estos casos, los encuentros con la IA pueden ser desconcertantes o desempoderadores, como cuando la IA recomienda a una persona en lugar de a otras para un ascenso o traslado deseable, o fomenta o promueve actitudes que desafían o superan la sabiduría imperante.

Para los directivos, el despliegue de la IA tendrá muchas ventajas. Las decisiones de la IA suelen ser tan precisas o más que las de los humanos y, con las salvaguardias adecuadas, pueden ser menos sesgadas. Del mismo modo, la IA puede ser más eficaz a la hora de distribuir recursos, predecir resultados y recomendar soluciones. De hecho, a medida que se generaliza la IA generativa, su capacidad para producir textos, imágenes, vídeos y códigos novedosos puede

incluso permitirle desempeñar con la misma eficacia que sus homólogos humanos funciones que suelen considerarse creativas (como la redacción de documentos y la creación de anuncios). Para el empresario que ofrece nuevos productos, el administrador que maneja nueva información y el desarrollador que crea una IA cada vez más potente, los avances en estas tecnologías pueden aumentar el sentido de agencia y elección.

Optimizar la distribución de los recursos y aumentar la precisión de la toma de decisiones es bueno para la sociedad, pero para el individuo, el sentido se deriva más a menudo de la autonomía y de la capacidad de explicar los resultados sobre la base de un conjunto de acciones y principios. Las explicaciones dan sentido y permiten un propósito; el reconocimiento público y la aplicación explícita de principios morales aportan justicia. Pero un algoritmo no ofrece razones basadas en la experiencia humana para explicar sus conclusiones al público en general. Algunas personas, sobre todo las que entienden de IA, pueden encontrar este mundo inteligible. Pero otras, más numerosas, pueden no entender por qué la IA hace lo que hace, lo cual disminuye su sentido de la autonomía y su capacidad de atribuir significado al mundo.

A medida que la IA transforma la naturaleza del trabajo, puede poner en peligro el sentido de identidad, realización y seguridad económica de muchas personas. Los más afectados por este cambio y la posible deslocalización probablemente ocuparán puestos de trabajo manuales y de mandos intermedios que requieren una formación específica, así como puestos profesionales que impliquen la revisión o interpretación de datos o la redacción de documentos en formularios estándar.¹ Aunque estos cambios pueden crear no solo nuevas eficiencias, sino también la necesidad de nuevos trabajadores, quienes experimentan la deslocalización, aunque sea a corto plazo, pueden obtener poco consuelo al saber que es un aspecto temporal de una transición que aumentará la calidad de vida general y la productividad económica de una sociedad. Algunos se verán liberados de la monotonía para centrarse en los elementos

más gratificantes de su trabajo. Otros se darán cuenta de que sus conocimientos ya no son punteros o ni siquiera necesarios.

Aunque estos retos son desalentadores, no carecen de precedentes. Revoluciones tecnológicas anteriores han desplazado o alterado el trabajo. Inventos como la hiladora mecánica desplazaron a los obreros e inspiraron el auge de los luditas, miembros de un movimiento político que pretendía prohibir —o, en su defecto, sabotear— las nuevas tecnologías para preservar sus antiguas formas de vida. La industrialización de la agricultura provocó migraciones masivas a las ciudades. La globalización alteró la fabricación y las cadenas de suministro, y ambas provocaron cambios, incluso disturbios, antes de que muchas sociedades acabaran asimilando los cambios para su mejora general. Sean cuales sean los efectos a largo plazo de la IA, a corto plazo la tecnología revolucionará determinados segmentos económicos, profesiones e identidades. Las sociedades deben estar preparadas para ofrecer a los desplazados no solo fuentes alternativas de ingresos, sino también fuentes alternativas de realización.

Toma de decisiones

En la era moderna, la reacción habitual ante un problema ha sido buscar una solución, a veces identificando a los actores humanos responsables de la deficiencia original. Este punto de vista ha asignado tanto la responsabilidad como la agencia a los seres humanos, y ambos han contribuido a nuestro sentido de quiénes somos. Ahora un nuevo actor entra en estas ecuaciones y puede disminuir nuestra sensación de que somos los principales pensadores y motores en una situación dada. En ocasiones, todos nosotros, tanto si creamos y controlamos la IA como si simplemente la utilizamos, interactuaremos con ella sin saberlo o recibiremos respuestas o resultados facilitados por la IA que no habíamos solicitado. En ocasiones, la IA invisible puede dotar al mundo de una simpatía mágica, como cuando las tiendas parecen anticiparse a nuestras visitas y caprichos. Otras veces, puede producir una

sensación kafkiana, como cuando las instituciones presentan decisiones que influyen en la vida —ofertas de empleo, decisiones sobre préstamos para coches y viviendas, o decisiones tomadas por empresas de seguridad o fuerzas de seguridad— que ningún ser humano puede explicar.

Estas tensiones —entre explicaciones razonadas y toma de decisiones opacas, entre individuos y grandes sistemas, entre personas con conocimientos técnicos y autoridad y personas sin ellos— no son nuevas. Lo nuevo es que su origen es otra inteligencia, no humana y a menudo inexplicable desde el punto de vista de la razón humana. Lo que también es nuevo es la omnipresencia y la escala de esta nueva inteligencia. Quienes carecen de conocimientos sobre la IA o de autoridad sobre ella pueden sentirse especialmente tentados a rechazarla. Frustrados por su aparente usurpación de su autonomía o temerosos de sus efectos adicionales, algunos pueden tratar de minimizar su uso de la IA y desconectarse de las redes sociales u otras plataformas de redes mediadas por IA, evitando su uso (al menos conscientemente) en su vida cotidiana.

Algunos segmentos de la sociedad pueden ir más allá, insistiendo en seguir siendo «fiscalistas» en lugar de «virtualistas». Como los amish y los menonitas, algunos individuos pueden rechazar la IA por completo, y plantarse firmemente en un mundo solo de fe y razón. Pero a medida que la IA se generalice, la desconexión se convertirá en un viaje cada vez más solitario. De hecho, incluso la posibilidad de desconexión puede resultar ilusoria: a medida que la sociedad se digitaliza cada vez más y la IA se integra cada vez más en gobiernos y productos, su alcance puede resultar casi ineludible.

Descubrimientos científicos

El desarrollo de la comprensión científica a menudo implica una brecha sustancial entre la teoría y el experimento, así como un considerable proceso de ensayo y error. Con los avances del aprendizaje automático, estamos empezando a ver un nuevo paradigma en el cual los modelos se derivan no de una comprensión

teórica, como ha sido tradicional, sino de una IA que extrae conclusiones basadas en resultados experimentales. Este enfoque requiere una pericia diferente de la que desarrolla modelos teóricos o modelos computacionales convencionales. Requiere no solo una comprensión profunda del problema, sino también saber qué datos, y qué representación de esos datos, serán útiles para entrenar un modelo de IA que lo resuelva. En el descubrimiento de la halicina, por ejemplo, la elección de qué compuestos, y qué atributos de esos compuestos, introducir en el modelo fue, por un lado, crucial y, por otro, fortuita.

El aumento de la importancia del aprendizaje automático para la comprensión científica ha supuesto un nuevo reto para la visión que tenemos de nosotros mismos y de nuestro papel o papeles en el mundo. Tradicionalmente, la ciencia ha sido una amalgama de pericia, intuición y perspicacia de origen humano. En la larga interacción entre teoría y experimentación, el ingenio humano impulsa todos los aspectos de la investigación científica. Pero la IA añade un concepto del mundo no humano —y divergente del humano— a la investigación, el descubrimiento y la comprensión científicas. El aprendizaje automático produce cada vez más resultados sorprendentes que impulsan nuevos modelos teóricos y experimentos. Al igual que los expertos en ajedrez han adoptado las sorprendentes estrategias de AlphaZero, interpretándolas como un reto para mejorar su propia comprensión del juego, los científicos de muchas disciplinas han empezado a hacer lo mismo. En las ciencias biológicas, químicas y físicas está surgiendo una asociación híbrida en la que la IA permite nuevos descubrimientos que los humanos, en respuesta, se esfuerzan por comprender y explicar.

Un ejemplo sorprendente de cómo la IA permite realizar descubrimientos de base amplia en las ciencias biológicas y químicas es el desarrollo de AlphaFold, que utilizó el aprendizaje por refuerzo para crear nuevos y potentes modelos de proteínas. Las proteínas son moléculas grandes y complejas que desempeñan un papel fundamental en la estructura, función y regulación de tejidos,

órganos y procesos de los sistemas biológicos. Una proteína está formada por cientos (o miles) de unidades más pequeñas llamadas aminoácidos, que se unen formando largas cadenas. Como existen veinte tipos diferentes de aminoácidos en la formación de proteínas, una forma habitual de representar una proteína es como una secuencia de cientos (o miles) de caracteres, en la que cada carácter procede de un «alfabeto» de veinte caracteres.

Aunque las secuencias de aminoácidos pueden ser muy útiles para estudiar las proteínas, no captan un aspecto fundamental de estas: la estructura tridimensional que forma la cadena de aminoácidos. Se puede pensar en las proteínas como formas complejas que deben encajar en un espacio tridimensional, como una cerradura y una llave, para que se produzcan determinados resultados biológicos o químicos, como la progresión de una enfermedad o su curación. En algunos casos, la estructura de una proteína puede medirse con métodos experimentales muy laboriosos, como la cristalografía. Pero en muchos casos, los métodos distorsionan o destruyen la proteína, lo que hace imposible medir la estructura. Por tanto, la capacidad de determinar la estructura tridimensional a partir de la secuencia de aminoácidos es fundamental. Desde la década de 1970, este reto se denomina plegamiento de proteínas.

Antes de 2016, no se había avanzado mucho en la mejora de la precisión del plegamiento de proteínas, hasta que un nuevo programa, AlphaFold, logró grandes avances. Como su nombre indica, AlphaFold se basó en el enfoque adoptado por los desarrolladores cuando enseñaron a AlphaZero a jugar al ajedrez. Al igual que AlphaZero, AlphaFold utiliza el aprendizaje por refuerzo para modelizar proteínas sin necesidad de conocimientos humanos, en este caso las estructuras proteicas conocidas en las que se basaban los métodos anteriores. AlphaFold ha duplicado con creces la precisión del plegamiento de proteínas, que ha pasado del 40 % al 85 %, lo cual ha permitido a biólogos y químicos de todo el mundo replantearse viejas cuestiones que no habían podido

responder y formular nuevas preguntas sobre la lucha contra patógenos en personas, animales y plantas.² Avances como AlphaFold —imposibles sin la IA— están superando los límites anteriores de la medición y la predicción. El resultado son cambios en la forma en que los científicos enfocan lo que pueden aprender para curar enfermedades, proteger el medio ambiente y resolver otros retos esenciales.

Educación y aprendizaje permanente

Alcanzar la mayoría de edad en presencia de la IA alterará nuestras relaciones, tanto entre nosotros como con nosotros mismos. Al igual que hoy existe una brecha entre los «nativos digitales» y las generaciones anteriores, también surgirá una brecha entre los «nativos de la IA» y las personas que les precedan. En el futuro, los niños podrán crecer con asistentes de IA, más avanzados que Alexa y Google Home, que serán muchas cosas a la vez: niñera, tutor, consejero, amigo. Un asistente de este tipo podrá enseñar a los niños prácticamente cualquier idioma o formarlos en cualquier materia, calibrando su estilo en función del rendimiento y los estilos de aprendizaje de cada alumno para sacar lo mejor de ellos. La IA puede servir de compañera de juegos cuando un niño se aburre y de monitora cuando sus padres están ausentes. A medida que se introduzca la educación proporcionada y adaptada por la IA, las capacidades del ser humano medio aumentarán y se pondrán a prueba.

La frontera entre los humanos y la IA es sorprendentemente porosa. Si los niños adquieren asistentes digitales a una edad temprana, se habituarán a ellos. Al mismo tiempo, los asistentes digitales evolucionarán con sus propietarios, interiorizando sus preferencias y prejuicios a medida que maduren. Un asistente digital encargado de maximizar la comodidad o satisfacción de un interlocutor humano mediante la personalización puede producir recomendaciones e información que se consideren esenciales

aunque el usuario humano no pueda explicar exactamente por qué son mejores que cualquier otro recurso alternativo.

Con el tiempo, los individuos pueden llegar a preferir sus asistentes digitales a los humanos, porque estos serán menos intuitivos de sus preferencias y más «desagradables» (aunque solo sea porque los humanos tienen personalidades y deseos que no están necesariamente en clave con otros individuos). Como resultado, nuestra dependencia de los demás, de las relaciones humanas, puede disminuir. ¿Qué será entonces de las inefables cualidades y lecciones de la infancia? ¿Cómo afectará a la percepción del mundo y a la socialización del niño la omnipresente compañía de una máquina que no siente ni experimenta emociones humanas (pero que puede imitarlas)? ¿Cómo influirá en la imaginación? ¿Cómo cambiará la naturaleza del juego? ¿Cómo alterará el proceso de hacer amigos o de encajar?

Podría decirse que la disponibilidad de información digital ya ha transformado la educación y la experiencia cultural de una generación. Ahora el mundo se embarca en otro gran experimento, en el que los niños crecerán con máquinas que, en muchos aspectos, actuarán como lo han hecho los profesores humanos durante generaciones, pero sin sensibilidad, perspicacia ni emoción humanas. Con el tiempo, los participantes en el experimento se preguntarán si sus experiencias se están alterando de un modo que no esperaban o aceptaban.

Los padres, alarmados por los efectos potencialmente inciertos de esta exposición en sus hijos, pueden oponerse. Al igual que los padres de hace una generación limitaban el tiempo de televisión y los padres de hoy limitan el tiempo de pantallas, los padres del futuro pueden limitar el tiempo de IA. Pero aquellos que quieran empujar a sus hijos al éxito, o que carezcan de la inclinación o la capacidad para sustituir la IA por un padre o tutor humano —o simplemente quieran satisfacer el deseo de sus hijos de tener amigos de IA— pueden aprobar la compañía de IA para sus hijos.

Así, los niños, que aprenden, evolucionan y son impresionables, pueden formarse una idea del mundo dialogando con las IA.

Lo irónico es que, al mismo tiempo que la digitalización aumenta la cantidad de información disponible, disminuye el espacio necesario para el pensamiento profundo y concentrado. El flujo casi constante de medios de comunicación aumenta el coste de la contemplación y, por tanto, disminuye su frecuencia. Los algoritmos promueven lo que capta la atención en respuesta al deseo humano de estimulación, y lo que capta la atención suele ser lo dramático, lo sorprendente y lo emocional. Una cuestión es si un individuo puede encontrar espacio en este entorno para pensar detenidamente. Otra es que las formas de comunicación ahora dominantes no favorecen el razonamiento moderado.

Nuevos intermediarios de información

Como decíamos en el capítulo 4, la IA configura cada vez más nuestro dominio informativo. Para informar y organizar la experiencia humana, se han creado intermediarios: organizaciones e instituciones que destilan información compleja, destacan lo que los individuos necesitan saber y difunden los resultados.³ A medida que las sociedades dividían cada vez más su trabajo físico, también dividían su trabajo mental, creando periódicos y revistas para informar a los ciudadanos en general y fundando universidades para educarlos específicamente. Desde entonces, la información ha sido agregada, destilada y difundida —y su significado definido— por dichas instituciones.

Ahora, en todos los ámbitos caracterizados por un intenso trabajo intelectual, desde las finanzas al derecho, la IA se está integrando en el proceso de aprendizaje. Pero los humanos no siempre podemos verificar que lo que presenta la IA es representativo; no siempre podemos explicar por qué aplicaciones como TikTok y YouTube promocionan unos vídeos sobre otros. Los editores y presentadores humanos, en cambio, pueden dar explicaciones

(exactas o no) de sus razones para seleccionar lo que presentan. Mientras la gente desee esa explicación, la era de la IA decepcionará a la mayoría de las personas que no entiendan los procesos y mecanismos de la tecnología.

Los efectos de la IA sobre el conocimiento humano son paradójicos. Por un lado, los intermediarios de la IA pueden navegar y analizar cuerpos de datos más vastos de lo que la mente humana sin ayuda podría haber contemplado antes. Por otro lado, este poder—la capacidad de trabajar con grandes volúmenes de datos—también puede acentuar las formas de manipulación y error. La IA es capaz de explotar las pasiones humanas con más eficacia que la propaganda tradicional. Al adaptarse a las preferencias e instintos individuales, la IA suscita las respuestas que su creador o usuario desea. Del mismo modo, el despliegue de intermediarios de IA también puede amplificar los sesgos inherentes, incluso si estos intermediarios de IA están técnicamente bajo control humano. La dinámica de la competencia en el mercado impulsa a las plataformas de redes sociales y a los motores de búsqueda a presentar la información que los usuarios consideran más atractiva. Como resultado, se prioriza la información que se cree que los usuarios quieren ver, y se distorsiona así una imagen representativa de la realidad. Del mismo modo que la tecnología aceleró la velocidad de producción y difusión de la información en los siglos XIX y XX, en esta era la información se está viendo alterada por el mapeo de la IA en los procesos de difusión.

Algunas personas buscarán filtros de información que no distorsionen, o al menos que distorsionen de forma transparente. Algunos equilibrarán filtro contra filtro, sopesando independientemente los resultados. Otros pueden renunciar por completo, prefiriendo la filtración por intermediarios humanos tradicionales. Sin embargo, cuando la mayoría de las personas de una sociedad acepten la intermediación de la IA, ya sea por defecto o como precio por alimentar las plataformas de redes, aquellos que persiguen formas tradicionales de indagación personal a través de la

investigación y la razón pueden verse incapaces de seguir el ritmo de los acontecimientos. Sin duda, verán progresivamente limitada su capacidad para darles forma.

Si la información y el entretenimiento se vuelven envolventes, personalizados y sintéticos —como las «noticias» clasificadas por IA que confirman las creencias arraigadas de la gente o las películas generadas por IA «protagonizadas» por actores fallecidos hace tiempo—, ¿tendrá una sociedad una comprensión común de su historia y su actualidad? ¿Tendrá una cultura común? Si se ordena a una IA que escanee un siglo de música o televisión y produzca «un éxito», ¿crea o simplemente ensambla? ¿Cómo se verán a sí mismos y cómo serán vistos por los demás los escritores, actores, artistas y otros creadores, cuyo trabajo se ha tratado tradicionalmente como un compromiso humano único con la realidad y la experiencia vivida?

Un nuevo futuro humano

La razón y la fe tradicionales persistirán en la era de la IA, pero su naturaleza y alcance se verán profundamente afectados por la introducción de una nueva y poderosa forma de lógica operada por máquinas. Puede que la identidad humana siga descansando en el pináculo de la inteligencia animada, pero la razón humana dejará de describir todo el alcance de la inteligencia que trabaja para comprender la realidad. Para dar sentido a nuestro lugar en este mundo, es posible que debamos pasar de la centralidad de la razón humana a la centralidad de la dignidad y la autonomía humanas.

La Ilustración se caracterizó por los intentos de definir la razón humana y comprenderla en relación con las épocas humanas anteriores y en contraste con ellas. Los filósofos políticos de la Ilustración —Hobbes, Locke, Rousseau y muchos otros— derivaron sus conceptos de los estados teóricos de la naturaleza, a partir de los cuales articularon puntos de vista sobre los atributos de los seres humanos y la estructura de la sociedad. A su vez, los líderes se preguntaban cómo se podría aunar y difundir objetivamente el

conocimiento humano para permitir un gobierno ilustrado y el florecimiento humano. Si no se realizan esfuerzos similares para comprender la naturaleza humana, será difícil mitigar las desorientaciones de la era de la inteligencia artificial.

Los cautelosos pueden tratar de restringir la IA, limitando su uso a funciones discretas y circunscribiendo cuándo, dónde y cómo se utiliza. Las sociedades o los individuos pueden reservarse el papel de principal y juez, relegando la IA a la posición de personal de apoyo. Sin embargo, la dinámica competitiva desafiará las limitaciones, de las que los dilemas de seguridad presentados en el capítulo anterior son la prueba más evidente. Salvo restricciones éticas o legales fundamentales, ¿qué empresa renunciaría a conocer las funcionalidades de la IA que ha utilizado un rival para ofrecer nuevos productos o servicios? Si la IA permite a un burócrata, arquitecto o inversor predecir resultados o conclusiones con facilidad, ¿en qué se basaría para no utilizarla? Dadas las presiones para su despliegue, las limitaciones a los usos de la IA que son, a primera vista, deseables, tendrán que formularse a nivel de toda la sociedad o a nivel internacional.

La IA puede asumir un papel protagonista en la exploración y gestión de los mundos físico y digital. En determinados ámbitos, los seres humanos pueden ceder ante la IA, prefiriendo sus procesos a las limitaciones de la mente humana. Esta deferencia podría llevar a muchos o incluso a la mayoría de los humanos a replegarse en mundos individuales, filtrados y personalizados. En este escenario, el poder de la IA —combinado con su prevalencia, invisibilidad y opacidad— planteará interrogantes sobre las perspectivas de las sociedades libres e incluso del libre albedrío.

En muchos ámbitos, la IA y el ser humano se convertirán en socios iguales en la empresa de la exploración. En consecuencia, la identidad humana llegará a reflejar la reconciliación con las nuevas relaciones, tanto con la IA como con la realidad. Las sociedades crearán esferas diferenciadas para el liderazgo humano. Al mismo tiempo, desarrollarán las estructuras sociales y los hábitos

necesarios para comprender e interactuar fructíferamente con la IA. Las sociedades deben crear la infraestructura intelectual y psicológica necesaria para interactuar con la IA y utilizar su inteligencia única en beneficio de los seres humanos en la mayor medida posible. La tecnología obligará a adaptarse en muchos —de hecho, en la mayoría— de los aspectos de la vida política y social.

En cada nuevo despliegue importante y discreto de la IA, será crucial establecer el equilibrio. Las sociedades y sus líderes tendrán que elegir cuándo se debe notificar a los individuos que están tratando con IA, así como qué poderes tienen en esas interacciones. En última instancia, estas decisiones pondrán de manifiesto una nueva identidad humana para la era de la IA.

Algunas sociedades e instituciones pueden adaptarse gradualmente. Otras, sin embargo, pueden encontrar que sus supuestos fundacionales entran en conflicto con la forma en que han llegado a percibir la realidad y a sí mismas. Como la IA facilita la educación y el acceso a la información, al tiempo que aumenta el potencial de amplificación y manipulación, estos conflictos pueden agravarse. Mejor informados, mejor equipados y con sus puntos de vista amplificados, los individuos pueden exigir más a sus gobiernos.

Surgen entonces varios principios. En primer lugar, para garantizar la autonomía humana, las decisiones gubernamentales fundamentales deben quedar al margen de las estructuras impregnadas de IA y limitarse a la administración y supervisión humanas. Los principios inherentes a nuestras sociedades prevén la resolución pacífica de los conflictos. En este proceso, el orden y la legitimidad están vinculados: el orden sin legitimidad es mera fuerza.

Garantizar la supervisión humana y la participación determinante en los elementos básicos del gobierno será esencial para mantener la legitimidad. En la administración de justicia, por ejemplo, las explicaciones y el razonamiento moral son elementos cruciales de la legitimidad, porque permiten a los participantes evaluar la imparcialidad de un tribunal e impugnar sus conclusiones si no se ajustan a los principios morales de la sociedad. Por lo tanto, en la

era de la IA, siempre que esté en juego una cuestión tan importante, las personas que tomen las decisiones deberán ser seres humanos cualificados, no anónimos, capaces de ofrecer razones para las decisiones tomadas.

Del mismo modo, la democracia debe conservar sus cualidades humanas. En el nivel más básico, esto significará proteger la integridad de las deliberaciones y elecciones democráticas. Una deliberación significativa requiere algo más que la oportunidad de hablar; también requiere la protección del discurso humano frente a la distorsión de la IA. La libertad de expresión debe continuar para los humanos, pero no extenderse a la IA. Como dijimos en el capítulo 4, la IA tiene la capacidad de generar, tanto en alta calidad como en gran volumen, desinformación como falsificaciones profundas, muy difíciles de distinguir de las grabaciones reales de vídeo y audio. Aunque el habla automatizada de la IA fue creada y desplegada a instancias de la gente, será importante desarrollar distinciones comprensibles entre ella y el habla humana genuina. Aunque será difícil regular la intermediación de la IA para evitar la promoción de la desinformación y la desinformación —falsedades creadas deliberadamente—, tal esfuerzo será crucial. En una democracia, la palabra permite a los ciudadanos compartir información relevante, participar deliberativamente en el proceso democrático y buscar la autorrealización a través de la producción de ficción, arte y poesía.⁴ Las declaraciones falsas generadas por la IA pueden aproximarse a la palabra humana, pero solo sirven para ahogarla o distorsionarla. Frenar la propagación de la IA que produce desinformación, por lo tanto, ayudaría a preservar el discurso que es vital para nuestro proceso deliberativo. ¿Clasifica uno un diálogo sobre IA entre dos figuras públicas que nunca se han visto como desinformación, entretenimiento o investigación política, o depende la respuesta del contexto o de los participantes? ¿Tiene un individuo derecho a no ser representado en una realidad simulada sin su permiso? Si se concede el permiso, ¿es la expresión sintética más genuina?

Cada sociedad debe determinar en primera instancia toda la gama de usos permitidos y no permitidos de la IA en diversos ámbitos. El acceso a ciertas IA potentes, como la AGI, deberá estar estrictamente vigilado para evitar usos indebidos. Como la construcción de la inteligencia artificial será probablemente tan costosa que solo unos pocos podrán hacerlo, el acceso puede estar intrínsecamente limitado. Ciertos límites pueden violar los conceptos de libre empresa y proceso democrático de una sociedad. Otros, como la necesidad de restringir el uso de la IA en la producción de armas biológicas, deberían acordarse fácilmente, pero requerirán la colaboración internacional.

En el momento de redactar este informe, la UE ha esbozado planes para regular la IA,⁵ tratando de equilibrar valores europeos como la privacidad y la libertad con la necesidad de desarrollo económico y el apoyo a las empresas europeas de IA. La normativa traza un camino intermedio entre el de China, donde el Estado está invirtiendo mucho en IA, incluso con fines de vigilancia, y el de Estados Unidos, donde la I+D en IA se ha dejado en gran medida en manos del sector privado. El objetivo de la UE es controlar la forma en que empresas y gobiernos utilizan los datos y la IA y facilitar la creación y el crecimiento de empresas europeas de IA. El marco regulador incluye evaluaciones de riesgo de diversos usos de la IA e impone límites o incluso prohibiciones al uso gubernamental de ciertas tecnologías consideradas de alto riesgo, como el reconocimiento facial (aunque este tiene usos beneficiosos, como la búsqueda de personas desaparecidas y la lucha contra la trata de seres humanos). Sin duda, habrá amplios debates y modificaciones del concepto inicial, pero su primera forma es un ejemplo de una sociedad que determina la gama de limitaciones de la IA que cree que le permitirán avanzar en su modo de vida y su futuro.

Con el tiempo, estos esfuerzos se institucionalizarán. En Estados Unidos, grupos académicos y órganos consultivos ya están empezando a examinar las relaciones entre los procesos y estructuras existentes y el auge de la inteligencia artificial. Se trata

de iniciativas académicas, como la del MIT sobre el futuro del trabajo,⁶ y gubernamentales, como la Comisión de Seguridad Nacional sobre Inteligencia Artificial.⁷ Se quedarán rezagadas con respecto a las sociedades que, gracias a la indagación, adaptan sus instituciones con antelación o, como analizamos en el capítulo siguiente, establecen instituciones completamente nuevas, reduciendo así las dislocaciones y maximizando los beneficios materiales e intelectuales que ofrece la asociación con la IA. A medida que se desarrolle la IA, la creación de este tipo de instituciones será crucial.

Percepción de la realidad y de la humanidad

La realidad explorada por la IA, o con su ayuda, puede resultar ser algo distinto de lo que los humanos habían imaginado. Puede tener patrones que nunca hemos discernido o que no podemos conceptualizar. Su estructura subyacente, penetrada por la IA, puede ser inexpressable solo con el lenguaje humano. Como ha observado uno de nuestros colegas a propósito de AlphaZero: «Ejemplos como este demuestran que hay formas de conocimiento que no están al alcance de la conciencia humana».⁸

Para trazar las fronteras del conocimiento contemporáneo, podemos encargar a la IA que explore reinos en los que no podemos entrar; tal vez regrese con patrones o predicciones que no acabamos de comprender. Los pronósticos de los filósofos gnósticos, sobre una realidad interior más allá de la experiencia humana ordinaria, pueden resultar de nuevo significativos. Puede que nos encontremos un paso más cerca del concepto de conocimiento puro, menos limitado por la estructura de nuestras mentes y los patrones del pensamiento humano convencional. No solo tendremos que redefinir nuestro papel como algo distinto del único conocedor de la realidad, sino que también deberemos redefinir la propia realidad que creíamos estar explorando. E incluso si la realidad no nos

desconcierta, la aparición de la IA puede alterar nuestro compromiso con ella y con los demás.

A medida que la IA se generaliza, algunas personas pueden considerar que la humanidad es más capaz que nunca de conocer y organizar su entorno. Otros pueden declarar que nuestra capacidad es menor de lo que creíamos. Estas redefiniciones de nosotros mismos y de la realidad en la que nos encontramos transformarán los supuestos básicos y, con ellos, los acuerdos sociales, económicos y políticos. El mundo medieval tenía su *imago dei*, sus modelos agrarios feudales, su veneración por la corona y su orientación hacia las alturas de la aguja de la catedral. La era de la razón tenía su *cogito ergo sum* y su búsqueda de nuevos horizontes y, con ello, nuevas afirmaciones de agencia dentro de las nociones de destino, tanto individuales como sociales. La era de la IA aún tiene que definir sus principios organizativos, sus conceptos morales o su sentido de las aspiraciones y limitaciones.

La revolución de la IA se producirá más rápidamente de lo que la mayoría de los humanos espera. A menos que desarrollemos nuevos conceptos para explicar, interpretar y organizar sus consiguientes transformaciones, no estaremos preparados para navegar por ella ni por sus implicaciones. Moral, filosófica, psicológica y prácticamente, en todos los sentidos, nos encontramos en el precipicio de una nueva época. Debemos recurrir a nuestros recursos más profundos —la razón, la fe, la tradición y la tecnología— para adaptar nuestra relación con la realidad de modo que siga siendo humana.

¹ David Autor, David Mindell y Elisabeth Reynolds, «The Work of the Future: Building Better Jobs in an Age of Intelligent Machines», MIT Task Force on the Work of the Future, November 17, 2020, <https://workofthefuture.mit.edu/research-post/the-work-of-the-future-building-better-jobs-in-an-age-of-intelligent-machines>.

² «AlphaFold: A Solution to a 50-Year-Old Grand Challenge in Biology», blog de DeepMind, November 30, 2020, <https://deepmind.com/blog/article/alphafold-a-solution-to-a-50-year-old-grand-challenge-in-biology>.

³ Véase Walter Lippmann, *Public Opinion* (New York: Harcourt, Brace and Company, 1922), 11.

⁴ Robert Post, «Participatory Democracy and Free Speech», *Virginia Law Review* 97, no. 3 (May 2011): 477-478.

⁵ European Commission, «A European Approach to Artificial Intelligence», <https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence>.

⁶ Autor, Mindell y Reynolds, «The Work of the Future».

⁷ Eric Schmidt, Robert Work, et al., Final Report: National Security Commission on Artificial Intelligence, March 2021, <https://www.nscai.gov/2021-final-report>.

⁸ Frank Wilczek, *Fundamentals: Ten Keys to Reality* (New York: Penguin Press, 2021), 205.

La IA y el futuro

Los cambios provocados por los avances de la imprenta en la Europa del siglo xv ofrecen una comparación histórica y filosófica con los retos de la era de la IA. En la Europa medieval se estimaba el conocimiento, pero los libros eran escasos. Los autores individuales producían literatura o recopilaciones enciclopédicas de hechos, leyendas y enseñanzas religiosas. Pero estos libros eran un tesoro reservado a unos pocos. La mayoría de las experiencias se vivían directamente y la mayor parte del conocimiento se transmitía oralmente.

En 1450, Johannes Gutenberg, orfebre de la ciudad alemana de Maguncia, utilizó dinero prestado para financiar la creación de una imprenta experimental. Su esfuerzo apenas tuvo éxito —su negocio naufragó y sus acreedores le demandaron—, pero en 1455 apareció la Biblia de Gutenberg, el primer libro impreso de Europa. En última instancia, su imprenta provocó una revolución que repercutió en todas las esferas de la vida occidental y, finalmente, mundial. Se calcula que en 1500 circulaban por Europa nueve millones de libros impresos, y que el precio de un libro individual había caído en picado. No solo la Biblia se distribuyó ampliamente en las lenguas de la vida cotidiana (en vez de en latín), sino que también empezaron a proliferar las obras de autores clásicos en los campos de la historia, la literatura, la gramática y la lógica.¹

Antes de la aparición del libro impreso, los europeos medievales accedían al conocimiento principalmente a través de las tradiciones comunitarias: participando en las cosechas y en los ciclos

estacionales, con su acumulación de sabiduría popular; practicando la fe y observando sus sacramentos en los lugares de culto; uniéndose a un gremio, aprendiendo sus técnicas y siendo admitidos en sus redes especializadas. Cuando se adquiría nueva información o surgían nuevas ideas (noticias del extranjero, un invento agrícola o mecánico innovador, interpretaciones teológicas novedosas), se transmitían oralmente a través de una comunidad o manualmente, mediante manuscritos copiados a mano.

Con la generalización de los libros impresos, la relación entre los individuos y el conocimiento cambió. La información y las ideas nuevas podían difundirse rápidamente por canales más variados. Los individuos podían buscar información útil para sus tareas específicas y enseñársela a sí mismos. Examinando los textos originales, podían poner a prueba las verdades aceptadas. Las personas con convicciones firmes y acceso a recursos modestos o a un mecenas, podían publicar sus ideas e interpretaciones. Los avances científicos y matemáticos podían transmitirse rápidamente, a escala continental. El intercambio de panfletos se convirtió en un método aceptado de disputa política, entrelazado con la disputa teológica. Las nuevas ideas se extendieron, a menudo derribando o remodelando fundamentalmente los órdenes establecidos, dando lugar a adaptaciones de la religión (la Reforma), revoluciones en la política (ajustando el concepto de soberanía nacional) y nuevos entendimientos en las ciencias (redefiniendo el concepto de realidad).

Hoy, una nueva época nos llama. En ella, una vez más, la tecnología transformará el conocimiento, el descubrimiento, la comunicación y el pensamiento individual. La inteligencia artificial no es humana. No espera, reza ni siente. Tampoco tiene conciencia ni capacidad de reflexión. Es una creación humana que refleja procesos diseñados por el hombre en máquinas creadas por el hombre. Sin embargo, en algunos casos, a una escala y velocidad impresionantes, produce resultados que se aproximan a los que, hasta ahora, solo se han alcanzado mediante la razón humana. A

veces, sus resultados asombran. En consecuencia, puede revelar aspectos de la realidad más dramáticos que los que jamás hayamos contemplado. Los individuos y las sociedades que recurran a la IA como aliada para ampliar sus habilidades o perseguir ideas pueden ser capaces de hazañas —científicas, médicas, militares, políticas y sociales— que eclipsen las de épocas precedentes. Sin embargo, una vez que las máquinas que se aproximan a la inteligencia humana se consideren claves para producir resultados mejores y más rápidos, la razón por sí sola puede llegar a parecer arcaica. Después de definir una época, el ejercicio de la razón humana individual puede ver alterado su significado.

La revolución de la imprenta en la Europa del siglo xv produjo nuevas ideas y discursos que alteraron y enriquecieron los modos de vida establecidos. La revolución de la IA está llamada a hacer algo parecido: acceder a nueva información, producir importantes avances científicos y económicos y, al hacerlo, transformar el mundo. Pero su impacto en el discurso será difícil de determinar. Al ayudar a la humanidad a navegar por la totalidad de la información digital, la IA abrirá perspectivas de conocimiento y comprensión sin precedentes. Alternativamente, su descubrimiento de patrones en masas de datos puede producir un conjunto de máximas que se acepten como ortodoxia en las plataformas de redes continentales y globales. Esto, a su vez, puede disminuir la capacidad humana de indagación escéptica que ha definido la época actual. Además, puede encauzar a determinadas sociedades y comunidades de plataformas de redes hacia ramas separadas y contradictorias de la realidad.

La IA puede mejorar o —si se utiliza mal— empeorar a la humanidad, pero el mero hecho de su existencia desafía y, en algunos casos, trasciende supuestos fundamentales. Hasta ahora, solo los humanos desarrollaban su comprensión de la realidad, una capacidad que definía nuestro lugar en el mundo y nuestra relación con él. A partir de ahí, elaboramos nuestras filosofías, diseñamos nuestros gobiernos y estrategias militares y desarrollamos nuestros

preceptos morales. Ahora, la IA ha revelado que la realidad puede conocerse de formas distintas, quizá más complejas, de lo que hasta ahora solo comprendían los humanos. A veces, sus logros pueden ser tan sorprendentes y desorientadores como los de los pensadores humanos más influyentes en su época de apogeo, y producir rayos de perspicacia y desafíos a los conceptos establecidos, todo lo cual exige un ajuste de cuentas. Con más frecuencia aún, la IA será invisible, se integrará en lo mundano y moldeará sutilmente nuestras experiencias de formas que nos resulten intuitivamente adecuadas.

Debemos reconocer que los logros de la IA, dentro de sus parámetros definidos, a veces están a la altura o incluso superan los que permiten los recursos humanos. Podemos consolarnos repitiendo que la IA es artificial, que no ha igualado o no puede igualar nuestra experiencia consciente de la realidad. Pero cuando nos encontramos con algunos de los logros de la IA —hazañas lógicas, avances técnicos, ideas estratégicas y gestión sofisticada de sistemas grandes y complejos— es evidente que estamos en presencia de otra experiencia de la realidad por parte de otra entidad sofisticada.

Gracias a la IA, se abren ante nosotros nuevos horizontes. Antes, los límites de nuestras mentes restringían nuestra capacidad para agregar y analizar datos, filtrar y procesar noticias y conversaciones, e interactuar socialmente en el ámbito digital. La IA nos permite navegar por estos ámbitos con mayor eficacia. Encuentra información e identifica tendencias que los algoritmos tradicionales no podían, o al menos no con la misma gracia y eficacia. Al hacerlo, no solo amplía la realidad física, sino que también permite expandir y organizar el floreciente mundo digital.

Pero, al mismo tiempo, la IA resta. Acelera dinámicas que erosionan la razón humana tal y como hemos llegado a entenderla: las redes sociales, que reducen el espacio para la reflexión, y la búsqueda en línea, que disminuye el impulso para la conceptualización. Los algoritmos anteriores a la IA eran buenos ofreciendo contenidos «adictivos» a los humanos. La IA es excelente

en ello. A medida que la lectura profunda y el análisis se contraen, también lo hacen las recompensas tradicionales por emprender estos procesos. A medida que aumenta el coste de no participar en el ámbito digital, crece su capacidad para influir en el pensamiento humano, para convencer, para dirigir, para desviar. Como consecuencia, el papel del ser humano en la revisión, comprobación y comprensión de la información disminuye. En su lugar, se amplía el papel de la IA.

Los románticos afirmaban que la emoción humana era una fuente de información válida y, de hecho, importante. Una experiencia subjetiva, sostenían, era en sí misma una forma de verdad. Los posmodernos llevaron la lógica de los románticos un paso más allá, cuestionando la posibilidad misma de discernir una realidad objetiva a través del filtro de la experiencia subjetiva. La IA llevará la cuestión mucho más lejos, pero con resultados paradójicos. Escaneará patrones profundos y revelará nuevos hechos objetivos: diagnósticos médicos, señales tempranas de desastres industriales o medioambientales, amenazas inminentes a la seguridad. Sin embargo, en el mundo de los medios de comunicación, la política, el discurso y el entretenimiento, la IA remodelará la información para ajustarla a nuestras preferencias, confirmando y profundizando potencialmente los prejuicios y, al hacerlo, reducirá el acceso y el acuerdo sobre una verdad objetiva. En la era de la IA, la razón humana se verá aumentada y disminuida a la vez.

A medida que la IA se entreteje en el tejido de la existencia cotidiana, amplía esa existencia y la transforma, la humanidad tendrá impulsos contradictorios. Enfrentados a tecnologías que van más allá de la comprensión de los no expertos, algunos pueden tener la tentación de tratar los pronunciamientos de la IA como juicios casi divinos. Tales impulsos, aunque equivocados, no carecen de sentido. En un mundo en el que una inteligencia que escapa a nuestra comprensión o control extrae conclusiones útiles pero ajenas a nosotros, ¿es absurdo someterse a sus juicios? Espoleados por esta lógica, puede producirse un reencantamiento del mundo, en el

que se confíe en las IA para pronunciamientos oraculares a los que algunos humanos se aferren sin cuestionarlos. Especialmente en el caso de la inteligencia artificial general (AGI, por sus siglas en inglés), los individuos pueden percibir una inteligencia divina, una forma sobrehumana de conocer el mundo e intuir sus estructuras y posibilidades.

Pero tal deferencia erosionaría el alcance y la escala de la razón humana y, por tanto, probablemente provocaría una reacción violenta. Del mismo modo que algunos optan por no utilizar las redes sociales, limitan el tiempo que los niños pasan frente a una pantalla y rechazan los alimentos modificados genéticamente, otros intentarán no participar en el «mundo de la IA» o limitar su exposición a los sistemas de IA para preservar el espacio de su razón. En las naciones liberales, estas opciones pueden ser posibles, al menos a nivel individual o familiar. Pero no serán gratuitas. Negarse a utilizar la IA significará no solo renunciar a comodidades como las recomendaciones automáticas de películas y las indicaciones para conducir, sino también dejar atrás vastos dominios de datos, plataformas de redes y avances en campos que van desde la atención sanitaria a las finanzas.

A nivel de la civilización, renunciar a la IA será inviable. Los líderes tendrán que enfrentarse a las implicaciones de la tecnología, sobre cuya aplicación tienen una gran responsabilidad.

La necesidad de una ética que comprenda e incluso guíe la era de la IA es primordial. Pero no puede confiarse a una sola disciplina o campo. Los informáticos y empresarios que están desarrollando la tecnología, los estrategas militares que pretenden desplegarla, los líderes políticos que se proponen darle forma y los filósofos y teólogos que pretenden sondear sus significados más profundos ven todos únicamente partes del cuadro. Todos deberían participar en un intercambio de puntos de vista no condicionado por ideas preconcebidas.

En todo momento, la humanidad tendrá tres opciones principales: limitar a la IA, asociarse con ella o diferir de ella. Estas opciones

definirán la aplicación de la IA a tareas o ámbitos específicos, y reflejarán dimensiones tanto filosóficas como prácticas. Por ejemplo, en emergencias aéreas o automovilísticas, ¿debería un copiloto de IA ceder el mando a un humano? ¿O al revés? En algunos casos, el rumbo evolucionará a medida que evolucionen las capacidades de la IA y los protocolos humanos para comprobar sus resultados. A veces, la deferencia y el respeto serán lo más adecuado: si una IA puede detectar el cáncer de mama en una mamografía antes y con más precisión que un humano, emplearla salvará vidas. A veces será mejor la colaboración, como en los vehículos autónomos que funcionan como los actuales pilotos automáticos de los aviones. Otras veces, sin embargo —como en contextos militares—, será fundamental establecer limitaciones estrictas, bien definidas y bien entendidas.

La IA transformará nuestro enfoque de lo que sabemos, cómo lo sabemos e incluso de lo que se puede saber. La era moderna ha valorado el conocimiento que las mentes humanas obtienen mediante la recopilación y el examen de datos y la deducción de ideas a través de observaciones. En esta era, el tipo ideal de verdad ha sido la proposición singular, verificable y demostrable mediante pruebas. Pero la era de la IA elevará un concepto de conocimiento que es el resultado de la asociación entre humanos y máquinas. Juntos, nosotros (los humanos) crearemos y ejecutaremos algoritmos (informáticos) que examinarán más datos de forma más rápida, más sistemática y con una lógica diferente a la que puede aplicar cualquier mente humana. A veces, el resultado será la revelación de propiedades del mundo que estaban más allá de nuestra concepción hasta que cooperamos con las máquinas.

La IA ya trasciende la percepción humana, en cierto sentido mediante la compresión cronológica o el «viaje en el tiempo»: habilitada por algoritmos y potencia de cálculo, analiza y aprende a través de procesos que a las mentes humanas les llevaría décadas o incluso siglos completar. En otros aspectos, el tiempo y la potencia de cálculo por sí solos no describen lo que hace la IA.

Inteligencia artificial general

¿Se aproximan los humanos y la IA a la misma realidad desde puntos de vista diferentes, con fuerzas complementarias? ¿O percibimos dos realidades diferentes que se solapan parcialmente: una que los humanos pueden elaborar mediante la razón y otra que la IA puede elaborar mediante algoritmos? Si es así, la IA percibe cosas que nosotros no percibimos ni podemos percibir, no solo porque no tenemos tiempo de razonar para llegar a ellas, sino también porque existen en un ámbito que nuestras mentes no pueden conceptualizar. La búsqueda humana por conocer el mundo en su totalidad se transformará, con el inquietante reconocimiento de que para alcanzar ciertos conocimientos puede que tengamos que confiar a la IA que los adquiera por nosotros y nos informe al respecto. En cualquier caso, a medida que la IA persiga objetivos cada vez más completos y amplios, los humanos la verán cada vez más como un «ser» que experimenta y conoce el mundo: una combinación de herramienta, mascota y mente.

Este rompecabezas no hará sino profundizarse a medida que los investigadores se acerquen o alcancen la AGI. Como escribimos en el capítulo 3, la AGI no se limitará a aprender y ejecutar tareas específicas, sino que, por definición, será capaz de aprender y ejecutar una amplia gama de tareas, muy parecidas a las que realizan los humanos. El desarrollo de AGI requerirá una inmensa potencia de cálculo, por lo tanto, probablemente solo sean creadas por unas pocas organizaciones bien financiadas. Aunque la AGI pueda distribuirse fácilmente, dadas sus capacidades, al igual que la IA actual será necesario restringir sus aplicaciones. Las limitaciones podrían imponerse permitiendo que solo las organizaciones autorizadas puedan utilizarla. Entonces, las preguntas serán: ¿quién controla la AGI? ¿Quién da acceso a ella? ¿Es posible la democracia en un mundo en el que unas pocas máquinas «geniales» son operadas por un reducido número de organizaciones? ¿Qué aspecto tiene, en estas circunstancias, la colaboración con la IA?

Si se produce la llegada de la inteligencia artificial, será un logro intelectual, científico y estratégico de primer orden. Pero no es necesario que se produzca para que la IA anuncie una revolución en los asuntos humanos.

El dinamismo y la capacidad de la IA para acciones y soluciones emergentes —es decir, inesperadas— la distinguen de las tecnologías anteriores. Sin regulación ni supervisión, las IA podrían desviarse de nuestras expectativas y, en consecuencia, de nuestras intenciones. La decisión de confinarla, asociarse con ella o aplazarla no la tomarán solo los humanos. En algunos casos, será dictada por la propia IA; en otros, por fuerzas auxiliares. Es posible que la humanidad emprenda una carrera hacia el abismo. A medida que la IA automatice procesos, permita a los humanos explorar grandes volúmenes de datos y organice y reorganice los mundos físico y social, las ventajas serán para los primeros. La competencia podría obligar a desplegar la inteligencia artificial sin tiempo suficiente para evaluar los riesgos, o sin tenerlos en cuenta.

Es esencial una ética de la IA. Cada decisión individual —limitar, asociarse o aplazar— puede tener o no consecuencias dramáticas, pero en conjunto se magnificarán. No pueden tomarse de forma aislada. Si la humanidad quiere dar forma al futuro, necesita ponerse de acuerdo sobre unos principios comunes que guíen cada elección. La acción colectiva será difícil, y a veces imposible, de lograr, pero las acciones individuales, sin una ética común que las guíe, solo magnificarán la inestabilidad.

Quienes diseñen, formen y se asocien con la IA podrán alcanzar objetivos a una escala y con un nivel de complejidad que, hasta ahora, han eludido a la humanidad: nuevos avances científicos, nuevas eficiencias económicas, nuevas formas de seguridad y nuevas dimensiones de supervisión y control social. Aquellos que no tengan esa capacidad de acción en el proceso de expansión de la IA y sus usos pueden llegar a sentirse observados, estudiados y manipulados por algo que no entienden y que no han diseñado ni elegido: una fuerza que opera con una opacidad que en muchas

sociedades no se tolera a los actores o instituciones humanas convencionales. Los diseñadores e implantadores de la IA deben estar preparados para responder a estas preocupaciones, sobre todo explicando a los no tecnólogos qué hace la IA, qué «sabe» y cómo lo hace.

Las cualidades dinámicas y emergentes de la IA generan ambigüedad en al menos dos aspectos. En primer lugar, la IA puede funcionar como esperamos pero generar resultados que no prevemos. Con esos resultados, puede llevar a la humanidad a lugares no previstos por sus creadores. Al igual que los estadistas de 1914 no supieron reconocer que la vieja lógica de la movilización militar, combinada con la nueva tecnología, llevaría a Europa a la guerra, el despliegue de la IA sin una cuidadosa consideración puede tener graves consecuencias. Estas pueden ser localizadas, como en el caso de un coche autónomo que toma una decisión que pone en peligro la vida, o trascendentales, como un conflicto militar importante. En segundo lugar, en algunas aplicaciones, la IA puede ser impredecible y sus acciones pueden resultar totalmente sorprendentes. Pensemos en AlphaZero que, en respuesta a la instrucción «gana al ajedrez», desarrolló un estilo de juego que, en los milenios de historia del juego, los humanos nunca habían concebido. Aunque los humanos pueden especificar cuidadosamente los objetivos de la IA, a medida que le damos más libertad, los caminos que toma para alcanzar sus objetivos pueden llegar a sorprendernos o incluso alarmarnos.

En consecuencia, los objetivos y autorizaciones de la IA deben diseñarse con cuidado, especialmente en campos en los que sus decisiones podrían ser letales. La IA no debe tratarse como algo automático. Tampoco debe permitírsele realizar acciones irrevocables sin supervisión humana, vigilancia o control directo. Creada por humanos, la IA debe ser supervisada por humanos. Pero en nuestra época, uno de los retos de la IA es que las habilidades y los recursos necesarios para crearla no están inevitablemente emparejados con la perspectiva filosófica para comprender sus implicaciones más

amplias. Muchos de sus creadores se preocupan sobre todo de las aplicaciones que pretenden hacer posibles y de los problemas que desean resolver: puede que no se paren a pensar si la solución puede producir una revolución de proporciones históricas o cómo puede afectar su tecnología a diversos grupos de personas. La era de la IA necesita su propio Descartes, su propio Kant, para explicar lo que se está creando y lo que significará para la humanidad.

El debate y la negociación razonados en los que participen gobiernos, universidades e innovadores del sector privado deberían tener como objetivo establecer límites a las acciones prácticas, como los que rigen las acciones de las personas y las organizaciones en la actualidad. La IA comparte atributos con algunos productos, servicios, tecnologías y entidades regulados, pero es distinta de ellos en aspectos vitales y carece de un marco conceptual y jurídico propio plenamente definido. Por ejemplo, las propiedades evolutivas y emergentes de la IA plantean retos regulatorios: qué y cómo opera en el mundo puede variar según los campos y evolucionar con el tiempo, y no siempre de forma predecible. La gobernanza de las personas se rige por una ética. La IA exige una ética propia que refleje no solo la naturaleza de la tecnología, sino también los retos que esta plantea.

Con frecuencia, no se aplican los principios vigentes. En la era de la fe, los tribunales determinaban la culpabilidad durante ordalías en las que el acusado se enfrentaba a un juicio por combate y se creía que Dios dictaba la victoria. En la era de la razón, la humanidad asignaba la culpabilidad según los preceptos de la razón, determinando la culpabilidad e imponiendo el castigo de acuerdo con nociones como la causalidad y la intención. Pero las IA no funcionan según la razón humana, ni tienen motivación, intención o autorreflexión humanas. En consecuencia, su introducción complica los principios de justicia existentes que se aplican a los humanos. Cuando un sistema autónomo que funciona sobre la base de sus propias percepciones y decisiones actúa, ¿es responsable su creador? ¿O el hecho de que la IA actúe la separa de su creador, al

menos en términos de culpabilidad? Si se recurre a la IA para vigilar indicios de delitos o para ayudar a juzgar la inocencia o la culpabilidad, ¿debe la IA ser capaz de «explicar» cómo ha llegado a sus conclusiones para que los funcionarios humanos las adopten?

En qué momento y en qué contextos de la evolución de la tecnología debe someterse a restricciones negociadas internacionalmente es otro tema esencial de debate. Si se intenta demasiado pronto, la tecnología puede quedar bloqueada, o puede haber incentivos para ocultar sus capacidades; si se retrasa demasiado, puede tener consecuencias perjudiciales, sobre todo en contextos militares. El reto se agrava por la dificultad de diseñar regímenes de verificación eficaces para una tecnología etérea, opaca y de fácil distribución. Los negociadores oficiales serán inevitablemente los gobiernos. Pero es necesario crear foros para tecnólogos, especialistas en ética, las empresas que crean y operan IA y otras personas ajenas a estos campos.

Para las sociedades, los dilemas que plantea la IA son profundos. Gran parte de nuestra vida social y política transcurre ahora en plataformas de redes habilitadas por IA. Este es especialmente el caso de las democracias, las cuales dependen de estos espacios de información para el debate y el discurso que conforman la opinión pública y confieren legitimidad. ¿Quién o qué instituciones deben definir el papel de la tecnología? ¿Quién debe regularla? ¿Qué papel deben desempeñar los individuos que utilizan la IA? ¿Las empresas que la producen? ¿Los gobiernos de las sociedades que la utilizan? Como parte de la respuesta a estas preguntas, deberíamos buscar modos de hacerla auditable, es decir, de hacer que sus procesos y conclusiones sean comprobables y corregibles. A su vez, la formulación de correcciones dependerá de la elaboración de principios que respondan a las formas de percepción y toma de decisiones de la IA. La moralidad, la volición e incluso la causalidad no se ajustan perfectamente a un mundo de IA autónomas. En la mayoría de los demás elementos de la sociedad, desde el transporte

a las finanzas o la medicina, se plantean versiones de estas cuestiones.

Pensemos en el impacto de la IA en las redes sociales. Gracias a las recientes innovaciones, estas plataformas se han convertido rápidamente en anfitrionas de aspectos vitales de nuestra vida en común. El hecho de que Twitter y Facebook destaquen, limiten o prohíban directamente contenidos o personas —todas ellas funciones que, como hemos visto en el capítulo 4, dependen de la IA— es una prueba de su poder. En particular, las naciones democráticas se verán cada vez más desafiadas por el uso de la IA en la promoción o eliminación unilateral, y a menudo opaca, de contenidos y conceptos. ¿Será posible conservar nuestra capacidad de acción a medida que nuestras vidas sociales y políticas se desplazan cada vez más hacia dominios comisariados por la IA, dominios en los que solo podemos navegar confiando en esa curaduría?

El uso de las IA para navegar por masas de información conlleva el reto de la distorsión, de que las IA promuevan el mundo que los humanos prefieren instintivamente. En este ámbito, resuenan nuestros prejuicios cognitivos, que las IA pueden magnificar fácilmente. Y con esas reverberaciones, con esa multiplicidad de opciones —unida al poder de seleccionar y filtrar—, prolifera la desinformación. Las empresas de medios sociales no publican noticias para promover una polarización política extrema y violenta. Pero es evidente que estos servicios no han maximizado el discurso ilustrado.

Inteligencia artificial, información libre y pensamiento independiente

¿Cuál debería ser entonces nuestra relación con la IA? ¿Deberíamos limitarla, potenciarla o colaborar en el gobierno de estos espacios? Que la distribución de cierta información —y, más aún, la desinformación deliberada— puede dañar, dividir e incitar es indiscutible. Son necesarios algunos límites. Sin embargo, la

prontitud con la que ahora se denuncia, combate y suprime la información perjudicial también debería incitar a la reflexión. En una sociedad libre, las definiciones de información perjudicial y desinformación no deben ser competencia exclusiva de las empresas. Pero si se confían a un panel o agencia gubernamental, ese organismo debería operar según normas públicas definidas y mediante procesos verificables para no ser objeto de explotación por parte de quienes detentan el poder. Si se confían a un algoritmo de IA, la función objetiva, el aprendizaje, las decisiones y las acciones de ese algoritmo deben ser claros y estar sujetos a revisión externa y al menos a alguna forma de apelación humana.

Naturalmente, las respuestas variarán de una sociedad a otra. Algunas pueden hacer hincapié en la libertad de expresión, posiblemente de forma diferente en función de su concepción relativa de la expresión individual, y limitar así el papel de la IA en la moderación de contenidos. Cada sociedad elegirá lo que valora, y esto puede dar lugar a complejas relaciones con los operadores de plataformas de redes transnacionales. La IA es porosa: aprende de los humanos, incluso cuando nosotros la diseñamos y le damos forma. Por tanto, no solo variarán las elecciones de cada sociedad, sino también la relación de cada sociedad con la IA, su percepción de la IA y los patrones que sus IA imitan y aprenden de los maestros humanos. No obstante, la búsqueda de los hechos y de la verdad no debe llevar a las sociedades a experimentar la vida a través de un filtro cuyos contornos no se pueden revelar ni probar. La experiencia espontánea de la realidad, en toda su contradicción y complejidad, es un aspecto importante de la condición humana, incluso cuando conduce a la ineficacia o al error.

IA y orden internacional

A escala mundial, son muchas las preguntas que exigen respuesta. ¿Cómo regular las plataformas de IA en redes sin provocar tensiones entre países preocupados por sus implicaciones para la seguridad? ¿Erosionarán estas plataformas de redes los conceptos tradicionales

de soberanía estatal? ¿Impondrán los cambios resultantes una polaridad en el mundo no conocida desde el colapso de la Unión Soviética? ¿Se opondrán las naciones pequeñas? ¿Tendrán éxito, o alguna esperanza de éxito, los esfuerzos por mediar en tales consecuencias?

A medida que aumenten las capacidades de la IA, será cada vez más importante y complicado definir el papel de la humanidad en su colaboración con ella. Se puede contemplar un mundo en el que los seres humanos dependan cada vez más de la IA en cuestiones de una magnitud cada vez mayor. En un mundo en el que un adversario despliegue con éxito la IA, ¿podrían los líderes que se defienden de él decidir responsablemente no desplegar la suya propia, aunque no estuvieran seguros de qué evolución presagiaría ese despliegue? Y si la IA poseyera una capacidad superior para recomendar un curso de acción, ¿podrían los responsables políticos negarse razonablemente, incluso si el curso de acción implicara un sacrificio de cierta magnitud? ¿Qué ser humano podría saber si el sacrificio era esencial para la victoria? Y si lo fuera, ¿querría realmente el responsable político negarse a ello? En otras palabras, puede que no tengamos más remedio que fomentar la IA. Pero también tenemos el deber de moldearla de forma que sea compatible con un futuro humano.

La imperfección es uno de los aspectos más perdurables de la experiencia humana, especialmente del liderazgo. A menudo, los responsables políticos se distraen con preocupaciones parroquiales. A veces, actúan basándose en suposiciones erróneas. Otras veces, actúan por pura emoción. Otras veces, la ideología distorsiona su visión. Cualesquiera que sean las estrategias que surjan para estructurar la asociación entre el ser humano y la IA, deben adaptarse. Si la IA muestra capacidades sobrehumanas en algunas áreas, su uso debe ser asimilable en contextos humanos imperfectos.

En el ámbito de la seguridad, los sistemas dotados de IA serán tan sensibles que los adversarios podrían intentar atacar antes de que los sistemas estén operativos. El resultado puede ser una

situación inherentemente desestabilizadora, comparable a la creada por las armas nucleares. Sin embargo, las armas nucleares se sitúan en un marco internacional de conceptos de seguridad y control de armas desarrollados durante décadas por gobiernos, científicos, estrategas y especialistas en ética, sujetos a perfeccionamiento, debate y negociación. La IA y las ciberarmas carecen de un marco comparable. De hecho, los gobiernos pueden ser reacios a reconocer su existencia. Las naciones —y probablemente las empresas tecnológicas— tienen que ponerse de acuerdo sobre cómo van a coexistir con la IA armada.

La difusión de la IA a través de las funciones de defensa de los gobiernos alterará el equilibrio internacional y los cálculos que lo han sustentado en gran medida en nuestra era. Las armas nucleares son costosas y, debido a su tamaño y estructura, difíciles de ocultar. La IA, por el contrario, funciona con ordenadores ampliamente disponibles. Debido a los conocimientos y recursos informáticos necesarios para entrenar modelos de aprendizaje automático, la creación de una IA requiere los recursos de grandes empresas o Estados-nación. Como la aplicación de la IA se lleva a cabo en ordenadores relativamente pequeños, la IA estará ampliamente disponible, incluso de formas no previstas. ¿Acabarán las armas basadas en IA al alcance de cualquiera que disponga de un ordenador portátil, una conexión a Internet y la capacidad de navegar por sus elementos oscuros? ¿Permitirán los gobiernos que agentes afiliados o no afiliados utilicen la IA para acosar a sus oponentes? ¿Crearán los terroristas ataques con IA? ¿Podrán atribuirlos (falsamente) a Estados u otros actores?

La diplomacia, que solía desarrollarse en un escenario organizado y predecible, tendrá amplios rangos, tanto de información como de operación. Los límites geográficos y lingüísticos seguirán diluyéndose. Los traductores de IA facilitarán el habla, sin el efecto moderador de la familiaridad cultural que conlleva el estudio lingüístico. Las plataformas de redes habilitadas por IA promoverán la comunicación a través de las fronteras. Por otra parte, la piratería

informática y la desinformación seguirán distorsionando la percepción y la evaluación. A medida que aumente la complejidad, será más difícil formular acuerdos aplicables y con resultados predecibles.

El injerto de la funcionalidad de la IA en las armas cibernéticas profundiza este dilema. La humanidad eludió la paradoja nuclear distinguiendo claramente entre las fuerzas convencionales —consideradas conciliables con la estrategia tradicional— y las armas nucleares, consideradas excepcionales. Mientras que las armas nucleares aplicaban la fuerza de forma contundente, las fuerzas convencionales eran discriminatorias. Pero las armas cibernéticas, capaces tanto de discriminación como de destrucción masiva, borran esta barrera. A medida que se les aplica la inteligencia artificial, estas armas se vuelven más imprevisibles y potencialmente más destructivas. Al mismo tiempo, como se mueven a través de las redes, estas armas desafían la atribución. También desafían la detección —a diferencia de las armas nucleares, pueden transportarse en memorias USB— y facilitan la difusión. Y en algunas formas, una vez desplegadas, pueden ser difíciles de controlar, sobre todo debido a la naturaleza dinámica y emergente de la IA.

Esta situación pone en tela de juicio la premisa de un orden mundial basado en normas. Además, da lugar a un imperativo: desarrollar un concepto de control de armamentos para la IA. En la era de la IA, la disuasión no funcionará a partir de preceptos históricos; no podrá hacerlo. Al principio de la era nuclear, las verdades desarrolladas en los debates entre destacados profesores (que tenían experiencia gubernamental) de Harvard, el MIT y Caltech condujeron a un marco conceptual para el control de armas nucleares que, a su vez, contribuyó a un régimen (y, en Estados Unidos y otros países, a organismos para aplicarlo). Aunque el pensamiento de los académicos fue importante, se llevó a cabo al margen del pensamiento del Pentágono sobre la guerra convencional: fue una adición, no una modificación. Pero los usos militares potenciales de la IA son más amplios que los de las armas

nucleares, y las divisiones entre ataque y defensa son, al menos actualmente, poco claras.

En un mundo de tal complejidad e impredecibilidad inherente, en el que las IA introducen otra posible fuente de percepciones erróneas y errores, tarde o temprano, las grandes potencias que poseen capacidades de alta tecnología tendrán que entablar un diálogo permanente. Dicho diálogo deberá centrarse en lo fundamental: evitar la catástrofe y, al hacerlo, sobrevivir.

La IA y otras tecnologías emergentes (como la computación cuántica) parecen acercar al ser humano al conocimiento de la realidad más allá de los confines de nuestra propia percepción. En última instancia, sin embargo, tal vez descubramos que incluso estas tecnologías tienen límites. Nuestro problema es que aún no hemos comprendido sus implicaciones filosóficas. Nos hacen avanzar, pero de forma automática y no consciente. La última vez que la conciencia humana sufrió un cambio significativo —la Ilustración—, la transformación se produjo porque la nueva tecnología engendró nuevas ideas filosóficas que, a su vez, fueron difundidas por la tecnología (en forma de imprenta). En nuestra época, la nueva tecnología se ha desarrollado, pero sigue necesitando una filosofía que la guíe.

La IA es una gran empresa con profundos beneficios potenciales. Los humanos la estamos desarrollando, pero ¿la emplearemos para mejorar nuestras vidas o para empeorarlas? Promete medicamentos más potentes, una atención sanitaria más eficaz y equitativa, prácticas medioambientales más sostenibles y otros avances. Al mismo tiempo, sin embargo, tiene la capacidad de distorsionar o, como mínimo, agravar la complejidad del consumo de información y la identificación de la verdad, y llevar a algunas personas a dejar que se atrofien sus capacidades de razonamiento y juicio independientes.

Otros países han hecho de la IA un proyecto nacional. Estados Unidos, como nación, aún no ha explorado sistemáticamente su alcance, estudiado sus implicaciones o iniciado el proceso de reconciliación con ella. Estados Unidos debe convertir todos estos

proyectos en prioridades nacionales. Este proceso requerirá la colaboración de personas con gran experiencia en diversos ámbitos, un proceso que se beneficiaría enormemente del liderazgo de un pequeño grupo de figuras respetadas de los más altos niveles de la administración, la empresa y el mundo académico, y que tal vez lo requiera.

Este grupo o comisión debería tener al menos dos funciones:

1. A nivel nacional, debe garantizar que el país siga siendo intelectual y estratégicamente competitivo en IA.
2. Tanto a nivel nacional como mundial, debe ser consciente, y concienciar, de las implicaciones culturales que produce la IA.

Además, el grupo debe estar preparado para colaborar con los grupos nacionales y subnacionales existentes.

Escribimos en medio de una gran empresa que abarca a todas las civilizaciones humanas, de hecho, a toda la especie humana. Sus iniciadores no lo concibieron necesariamente así; su motivación era resolver problemas, no reflexionar o remodelar la condición humana. La tecnología, la estrategia y la filosofía deben alinearse, no sea que una supere a las otras. ¿Qué debemos proteger de la sociedad tradicional? ¿Y qué de la sociedad tradicional debemos arriesgar para conseguir una superior? ¿Cómo pueden integrarse las cualidades emergentes de la IA en los conceptos tradicionales de normas sociales y equilibrio internacional? ¿Qué otras preguntas deberíamos intentar responder cuando, para la situación en la que nos encontramos, no tenemos experiencia ni intuición?

Por último, se plantea una pregunta «meta»: ¿pueden los seres humanos satisfacer la necesidad de filosofía asistidos por las IA, que interpretan y, por tanto, comprenden el mundo de forma diferente? ¿Nuestro destino es que los humanos no comprendan del todo a las máquinas, pero hagan las paces con ellas y, al hacerlo, cambien el mundo?

Immanuel Kant abría el prefacio de su *Crítica de la razón pura* con una observación:

La razón humana tiene, en una especie de sus conocimientos, el destino particular de verse acosada por preguntas que no puede descartar, porque le son dadas como problemas por la naturaleza de la razón misma, pero que tampoco puede responder, porque sobrepasan toda la capacidad de la razón humana.²

En los siglos transcurridos desde entonces, la humanidad ha profundizado en estas cuestiones, algunas de las cuales se refieren a la naturaleza de la mente, la razón y la propia realidad. Y la humanidad ha hecho grandes avances. También se ha encontrado con muchas de las limitaciones que planteaba Kant: un reino de preguntas que no puede responder, de hechos que no puede conocer plenamente.

La llegada de la IA, con su capacidad para aprender y procesar la información de un modo que la razón humana por sí sola no puede, probablemente suponga un avance en cuestiones que han demostrado estar más allá de nuestra capacidad de respuesta. Pero el éxito producirá nuevas preguntas, algunas de las cuales hemos intentado articular en este libro. La inteligencia humana y la inteligencia artificial se están encontrando, aplicándose a actividades a escala nacional, continental e incluso mundial. Comprender esta transición y desarrollar una ética que la guíe exigirá el compromiso y la perspicacia de muchos elementos de la sociedad: científicos y estrategas, estadistas y filósofos, clérigos y directores ejecutivos. Este compromiso debe asumirse dentro de las naciones y entre ellas. Ahora es el momento de definir tanto nuestra asociación con la inteligencia artificial como la realidad que resultará de ella.

¹ J. M. Roberts: *Historia del mundo. De la prehistoria a nuestros días*. Editorial Debate, Barcelona, 2010.

² Immanuel Kant: *Crítica de la razón pura*, trad. Pedro Ribas. Taurus, Madrid, 2013.

Una nueva realidad

En los meses transcurridos desde la publicación original de este texto, una serie de evoluciones descritas en estas páginas han continuado a buen ritmo y han surgido *nuevos* fenómenos. Volvemos sobre nuestro tema para trazar brevemente estos desarrollos recientes.

Desde las redes eléctricas hasta las cadenas de suministro, pasando por el descubrimiento de fármacos,¹ la IA ha demostrado su potencial revolucionario. En los próximos años, la IA acelerará los avances en la mitigación del cambio climático, transformará la agricultura y revolucionará la medicina.

Estos cambios pueden llegar antes de lo que pensábamos. Postulamos que, para 2040, la inteligencia artificial podría ser quizás un millón de veces más potente de lo que era en 2021, siguiendo la «Ley de Moore» que predice la duplicación de la potencia de procesamiento informático cada dos años. Aunque el aumento de la potencia de la IA es más difícil de cuantificar que el de la potencia informática, parece que el crecimiento es aún más rápido. Por ejemplo, la potencia de los grandes modelos lingüísticos, las redes neuronales en las que se basa gran parte del procesamiento del lenguaje natural actual, está creciendo aún más rápido: se ha triplicado en menos de dos años. El Turing Megatron² de Microsoft, lanzado a finales de 2021, y el PaLM de Google,³ a principios de 2022, tienen cada uno más de 525 mil millones de parámetros, frente a los 175 mil millones del GPT-3 de OpenAI, en una amplia

gama de tareas lingüísticas. OpenAI también está trabajando en su próxima versión de GPT, para continuar así la carrera.

Modelos como el RETRO, de DeepMind, y el GLIDE, de OpenAI, han mejorado tanto su eficiencia como su capacidad; pueden hacer más con el mismo número de parámetros del modelo pero utilizando a menudo más datos de entrenamiento.⁴ A medida que estos modelos añaden nodos, capas y conexiones, pueden reconocer y utilizar más relaciones y patrones. Juntos, estos avances sugieren una aceleración hacia máquinas con habilidades lingüísticas más polifacéticas y sofisticadas y con mucha mayor capacidad de aprendizaje.

Estos sistemas tienen la capacidad de generar y editar el lenguaje en una variedad de medios. Estimulada por unas pocas palabras, la IA puede completar automáticamente un artículo o una historia prediciendo lo que un autor humano habría escrito a continuación, traducir entre idiomas, resolver problemas de palabras y generar explicaciones de «cadena de pensamiento». Más allá del lenguaje humano, pueden predecir terminaciones de fragmentos de código informático, proporcionar posibles correcciones de errores en el código informático o incluso generar traducciones de una descripción de texto humano a código informático. A modo de ilustración, la versión de 540 mil millones de parámetros de PaLM es capaz de proporcionar explicaciones de un chiste como el siguiente: «¿Has visto que Google acaba de contratar a una elocuente ballena para su equipo de TPU? Les ha enseñado a comunicarse entre diferentes pods (manadas)». PaLM ofrece la explicación: «Las TPU son un tipo de chip informático que Google utiliza para el aprendizaje profundo. Un "pod" es un grupo de TPU. Un "pod" es también un grupo de ballenas. El chiste es que la ballena es capaz de comunicarse entre dos grupos de ballenas, pero el orador está diciendo que la ballena es capaz de comunicarse entre dos grupos de TPU».

El rendimiento de los modelos lingüísticos crece de un modo cada vez más impresionante en una gama más amplia de tareas, pero al mismo tiempo sigue siendo importante recordar la advertencia del

capítulo 1, en el que GPT-3 describe sus propias capacidades como «un modelo lingüístico y no una máquina de razonamiento». Los modelos lingüísticos codifican lo que se refleja en el texto humano, sin una comprensión más profunda, aunque a veces puedan proyectar la apariencia de esa comprensión más profunda.

Durante el último año, el aprendizaje automático ha ido ampliando su visión, por así decirlo. Como predijo el científico jefe de OpenAI hace dos años, los modelos lingüísticos «han empezado a ser conscientes del mundo visual».⁵ La multimodalidad, la capacidad de los modelos lingüísticos entrenados en texto para procesar y generar medios sonoros y visuales, es un floreciente campo de exploración. El ejemplo actual más destacado es DALL-E 2, de OpenAI, anunciado a principios de 2022.⁶ DALL-E 2 puede crear imágenes de calidad fotográfica o artística a partir de descripciones de texto arbitrarias. Por ejemplo, crear imágenes realistas de «gatos jugando al ajedrez», o «un salón lleno de arena, arena en el suelo, con un piano en la habitación».⁷ Se trata de imágenes originales, aunque originalidad no siempre significa creatividad. De hecho, podría decirse que gran parte de la creatividad procede de la experimentación de la gente con las descripciones de texto, y de la colaboración entre la creatividad humana y un socio muy sofisticado para traducir el texto en imágenes.

Estos modelos, y otros similares, demuestran que los ordenadores van adquiriendo nuevas habilidades. Los modelos de solo texto se están convirtiendo en modelos de texto e imágenes. Y se están combinando varios modelos para obtener mejoras. Estos avances confirman nuestra intuición, comentada en el capítulo 3, de que la acumulación de habilidades podría ser el camino hacia una inteligencia más general. Por ejemplo, permitir que un único sistema realice tanto las tareas verbales como las visuales que hemos descrito.

Ese único sistema podría adoptar la forma de un compañero digital de IA. Este acompañante actuaría como un asistente: proporcionaría información en respuesta a preguntas, resumiría

largos artículos y libros y revisaría y comentaría los medios sociales. Podría sugerir o completar las compras habituales, enlazando con minoristas en línea o encargando la compra automáticamente, sugerir películas o podcasts, e incluso crear mensajes y vídeos humorísticos para compartir con los amigos. Muchas de estas capacidades ya existen en sistemas independientes, pero el compañero digital las reuniría todas. Al igual que un asistente humano, este acompañante se volvería más útil con el entrenamiento.

Los continuos avances en los modelos lingüísticos de IA, los modelos visuales y otros modelos más especializados darán lugar a desarrollos de inteligencia artificial que pasarán de actuar como nuestros asistentes a convertirse también en nuestros interlocutores y colaboradores. En los próximos años, es posible que podamos crear «gemelos IA» de nosotros mismos, segundos «yo» que hablen y se comporten como nosotros. Y esos gemelos nos sobrevivirán.

Estas tendencias amplificarán el poder de las plataformas de redes globales. A medida que las herramientas de IA se conviertan en narradoras y cocreadoras de nuestros vídeos, los contenidos en línea serán más entretenidos y prolíficos, pero también harán cada vez más difícil distinguir la expresión humana de la de las máquinas. Ya estamos experimentando un diluvio de contenidos en línea, con la posibilidad muy real de que el material creado por la IA ya esté, o lo haga muy pronto, ahogando la expresión humana directa. Esto puede exigir una definición de la libertad de expresión en la era de la IA que distinga entre la originada por humanos frente a la procedente de máquinas.

Los avances de la IA para el descubrimiento científico también han seguido acelerándose. En el verano de 2021, DeepMind lanzó Alpha Fold2, el sucesor de AlphaFold para predecir la estructura tridimensional de las proteínas a partir de su secuencia de aminoácidos, del que ya hablamos en el capítulo 6. AlphaFold2 y el trabajo del laboratorio Baker de la Universidad de Washington fueron nombrados por Science «Avance del año 2021»,⁸ y ha generado una

nueva base de datos de DeepMind de estructuras de proteínas que contiene casi un millón de proteínas a partir de la primavera de 2022, con planes de hacerla crecer hasta casi 100 millones (cientos de veces más que el número de estructuras que se han determinado experimentalmente). Estos avances están acelerando una amplia gama de investigaciones biológicas, incluso en el importante campo de las interacciones proteína-proteína para problemas como las vacunas contra el virus SARS-CoV-2.

La creciente sofisticación del papel de la IA en los descubrimientos científicos nos está conduciendo a un mundo en el que la IA puede estar haciendo contribuciones sustanciales a las invenciones, a un nivel que si fuera asumido por las personas las convertiría en coinventoras. La ley de patentes en EE. UU. y en la mayoría de las naciones solo reconoce a los inventores humanos, lo cual podría conducir con el tiempo a situaciones en las que existan invenciones sin inventores (o quizá con personas que desempeñen un papel de coinventor minimis).

Los gobiernos nacionales han reconocido la amenaza de la IA para el lenguaje: Hungría ha encargado su propio gran modelo lingüístico para que el húngaro no se quede automáticamente obsoleto en el ámbito digital.⁹ Los gobiernos también han empezado a lidiar con la dilución de la identidad comunal y nacional por el impacto de las redes digitales. Algunos Estados, como China y Rusia, han sido más agresivos en este esfuerzo que otros. El equilibrio entre los riesgos de anarquía y los de opresión se está alcanzando de forma diferente en los distintos lugares. La esperanza de que existan normas mundiales para controlar la desinformación y otros efectos secundarios de las plataformas de redes mundiales es cada vez menor.

Nuestras sociedades y alianzas también están abordando con renovado vigor el problema de la regulación, ante los importantes desafíos que la IA plantea a la seguridad, los conflictos y el orden mundial. No se trata de retos teóricos. Los avances en la capacidad de la IA para fabricar y manipular el ADN están creando nuevos

peligros en la guerra química y biológica.¹⁰ En concreto, surge la posibilidad de adaptar un arma biológica a poblaciones que comparten rasgos genéticos específicos. Ya se han desplegado armas autónomas en el campo de batalla.¹¹ El Departamento de Defensa entrena a sus «agentes IA» en el «gimnasio OpenAI». ¹² Cómo garantizar que los sistemas automáticos calibren adecuadamente la proporcionalidad y prevengan los errores sin perder el tiempo —masivamente comprimido por la automatización— es un reto permanente. Más recientemente, la tecnología de ClearviewAI, una empresa que «raspa» o recopila imágenes en línea con fines de reconocimiento facial, ha sido utilizada por Ucrania para identificar a camaradas muertos y a soldados rusos vivos.¹³ No sabemos con qué precisión.

Los reguladores están dando pasos hacia el control de los efectos de la IA en los ámbitos civil y militar: Han proliferado los proyectos de ley relacionados con la IA en Estados Unidos, Europa y China, aunque con enfoques bastante diferentes y con los consiguientes beneficios y riesgos potenciales para la regulación. La OTAN publicó una estrategia de defensa contra la IA justo después de la publicación de este libro. Sus principios para un uso responsable son ambiciosos, pero fundamentalmente correctos: las aplicaciones de IA en defensa deben ser cuidadosamente construidas, responsables, comprensibles, imparciales y limitadas a sus usos y funciones previstos.¹⁴

Sin embargo, siguen existiendo riesgos inherentes. A medida que crecen las capacidades de los modelos, su «gama completa de usos» se hace casi imposible de predecir con certeza. Probar los puntos débiles de un modelo debe hacerse en masa, a escala, y no solo por los desarrolladores del modelo. Y, si uno o unos pocos modelos llegan a dominar —como esperamos que ocurra—, nos enfrentamos al peligro de una «monocultura tecnológica» que corre el riesgo de multiplicar cualquier debilidad, sesgo o limitación de un pequeño número de modelos en una gran variedad de ámbitos. Incluso sin tales riesgos, la mera reducción de la diversidad y variabilidad de los

resultados que se deriva de contar con un pequeño número de modelos hace que la naturaleza de estos sea de crucial importancia. Imagine sustituir las decisiones independientes de millones o incluso miles de millones de personas por un puñado de personas, o en este caso un puñado de modelos de IA.

Estos ordenadores funcionarán con una inteligencia claramente no humana o, mejor dicho, con múltiples inteligencias no humanas. En los últimos meses, por ejemplo, la inteligencia del ordenador se ha aventurado cada vez más en las matemáticas y ha detectado patrones nunca vistos por la mente humana.¹⁵ Careceremos de una comprensión real de estas inteligencias, incluso cuando imiten el comportamiento humano. Durante un tiempo, nos limitaremos a observar sus patrones y a tratar de elaborar leyes que regulen eficazmente sus efectos.

Con el tiempo, sin embargo, si queremos asociarnos con estos sistemas, tendremos que desarrollar una «teoría de la mente» para algo muy nuevo.

Los retos de la interacción entre humanos y máquinas están aumentando rápidamente. Todo el conocimiento humano ha existido como en una rueda, cada segmento conectando con el siguiente, de la filosofía a la ciencia, a la política, a la historia, a la literatura, al arte y vuelta a empezar.¹⁶ ¿Cómo se aplica esto a los cambios revolucionarios provocados por la IA? En la primera fase, es probable que la inteligencia artificial se sume a nuestra rueda del conocimiento tal como es, presionando las fronteras de cada materia hacia nuevos terrenos. Esto ampliará la capacidad de los humanos y acelerará el aprendizaje. Lo hará en gran medida dentro de los marcos conceptuales conocidos para la adquisición de conocimientos. Se trata de una evolución a la que, con tiempo suficiente, podremos adaptarnos, siempre que no nos volvamos excesivamente dependientes del ordenador. De hecho, la delegación de asuntos mundanos a la IA puede liberar tiempo para el pensamiento conceptual y una mayor expansión del conocimiento humano.

Sin embargo, en una probable segunda fase, los ordenadores pueden añadir nuevos tipos de conocimiento, y alcanzar niveles imposibles de lograr por los seres humanos. La dependencia de conocimientos que solo pueden obtenerse a través de las máquinas traslada a la humanidad a una nueva realidad.

El desarrollo de esta realidad sería distinto al de la Ilustración. Entonces, cada expansión del conocimiento humano se basaba en logros anteriores de la razón y el intelecto. Como observamos en el capítulo 2, Montesquieu caracterizó el progreso del pensamiento de la Ilustración como «una cadena en la que cada idea precede a otra y sigue a otra». Al carecer de una progresión similar, la IA proporciona respuestas sin nociones lógicas humanas sobre cómo una idea sigue a otra. Se trata de un modo de pensamiento y conocimiento totalmente diferente. Puede resultar desconcertante para los humanos que confían en la razón humana para tener un sentido de significado y agencia.

Esta nueva inteligencia y su forma diferente de lógica cambiarán la percepción humana de la realidad, como la nueva inteligencia de AlphaZero está cambiando la percepción humana del ajedrez. Un fenómeno así requiere su propia filosofía, que habrá que desarrollar. Si no lo hacemos, podríamos encontrarnos desechando por completo la vieja rueda del conocimiento. Embelesados por máquinas que aparecen como nuestros amigos, temerosos de bloquear su velocidad sobrehumana e incapaces de explicar sus nuevas conclusiones, los humanos pueden desarrollar una reverencia por los ordenadores cercana al misticismo. Los papeles de la historia, la moralidad, la justicia y el juicio humano en un mundo así no están claros.

La inteligencia artificial conducirá a los seres humanos a reinos que no podemos alcanzar únicamente con la razón humana, ni ahora, ni quizá nunca. Sus logros técnicos en materia de salud y economía prometen hacer de la era de IA una era de abundancia. Mientras celebramos ese potencial, reconocemos que está surgiendo

una nueva realidad. A medida que aumenta el desafío, nuestra respuesta debe estar a su altura.

¹ Nathan Benaich and Ian Hogarth, «State of AI Report», October 12, 2021, <https://www.stateof.ai/2021-report-launch.html>.

² Ali Alvi and Paresh Kharya, «Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B...», October 11, 2021, <https://www.microsoft.com/en-us/research/blog/using-deepspeed-and-megatron-to-train-megatron-turing-nlg-530b-the-worlds-largest-and-most-powerful-generative-language-model/>.

³ Sharan Narang and Aakanksha Chowdhery, «Pathways Language Model (PaLM)...», April 4, 2022, <https://ai.googleblog.com/2022/04/pathways-language-model-palm-scaling-to.html>.

⁴ Will Douglas Heaven, «DeepMind Says Its New Language Model Can Beat Others 25 Times Its Size», MIT Technology Review, December 8, 2021, <https://www.technologyreview.com/2021/12/08/1041557/deepmind-language-model-beat-others-25-times-size-gpt-3-megatron/>.

⁵ *The Batch*, Issue 72, December 30, 2021, <https://read.deeplearning.ai/the-batch/issue-72/>.

⁶ <https://openai.com/dall-e-2/>.

⁷ Cade Metz, «Meet DALL-E, the A.I. That Draws Anything at Your Command», *New York Times*, April 6, 2022, <https://www.nytimes.com/2022/04/06/technology/openai-images-dall-e.html>.

⁸ Robert Service, «Protein structures for all», *Science*, December 16, 2021, <https://www.science.org/content/article/breakthrough-2021>.

⁹ David F. Carr, «Hungarian Gov Teams Up With Eastern European Bank to Develop AI Supercomputer», December 9, 2021, <https://venturebeat.com/2021/12/09/hungarian-gov-teams-up-with-eastern-european-bank-to-develop-ai-supercomputer/>.

¹⁰ David Gisselsson, «Next-Generation Biowarfare: Small in Scale, Sensational in Nature?», *Health Security*, April 22, 2022, <https://www.liebertpub.com/doi/full/10.1089/hs.2021.0165>.

¹¹ Nathaniel Allen and Marian «Ify» Okpali, «Artificial Intelligence Creeps on to the African Battlefield», *TechStream*, February 2, 2022, <https://www.brookings.edu/techstream/artificial-intelligence-creeps-on-to-the-african-battlefield/>.

¹² Department of Defense Small Business Innovation Research (SBIR) Program, https://media.defense.gov/2022/Apr/15/2002977654/-1/-1/1/DOD_22.2_FULL.PDF.

¹³ Paresh Dave and Jeffrey Dastin, «Exclusive: Ukraine Has Started Using Clearview AI's Facial Recognition During War», Reuters, March 14, 2022, <https://www.reuters.com/technology/exclusive-ukraine-has-started-using-clearviewais-facial-recognition-during-war-2022-03-13/>.

¹⁴ Summary of the NATO Artificial Intelligence Strategy, October 22, 2021, https://www.nato.int/cps/en/natohg/official_texts_1876r7.htm.

¹⁵ Kyle Wiggers, «DeepMind Claims AI Has Aided New Discoveries and Insights in Mathematics», December, 2021, <https://venturebeat.com/2021/12/01/deepmind-claims-aihas-aided-new-discoveries-and-insights-in-mathematics/>.

¹⁶ Como enseñó a sus alumnos el difunto diplomático y profesor Charles Hill.

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

Responsable editorial: Eugenio Tuya Feijoó

Edición en formato digital: 2023

THE AGE OF A.I.

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

All rights reserved

© EDICIONES ANAYA MULTIMEDIA (GRUPO ANAYA, S. A.), 2023

Calle Valentín Beato, 21

28037 Madrid

www.anayamultimedia.es

ISBN ebook: 978-84-415-4892-3

Está prohibida la reproducción total o parcial de este libro electrónico, su transmisión, su descarga, su descompilación, su tratamiento informático, su almacenamiento o introducción en cualquier sistema de repositorio y recuperación, en cualquier forma o por cualquier medio, ya sea electrónico, mecánico, conocido o por inventar, sin el permiso expreso escrito de los titulares del Copyright.