¿QUÉ SABEMOS DE?

La ética de la inteligencia artificial

Sara Degli-Esposti

CSIC

CATARATA

Sara Degli-Esposti

Es investigadora científica en Ética e Inteligencia Artificial en el Instituto de Filosofía del Consejo Superior de Investigaciones Científicas. Es coinvestigadora principal del proyecto “Ingeniería de valores en sistemas de IA – VAE” y Ethics Advisor del proyecto H2020 NEST. Su investigación se centra en inteligencia artificial fiable y confiable, riesgos de privacidad de las tecnologías digitales y cumplimiento normativo con las leyes de protección de datos en el contexto del capitalismo de la vigilancia.


CSIC


CATARATA

La ética de la inteligencia artificial

Sara Degli-Esposti

CSIC

CATARATA

ISBN (CSIC): 978-84-00-11199-1
ISBN electrónico (CSIC): 978-84-00-11200-4
ISBN (Catarata): 978-84-1352-841-0
ISBN electrónico (Catarata): 978-84-1352-842-7
NIPO: 833-23-128-4
NIPO electrónico: 833-23-129-X
Depósito legal: M-30.725-2023
Thema: PDZ/PDR/UX

Índice

Introducción

Nuestra migración voluntaria a entornos digitales para desarrollar cualquier tipo de actividad, desde lo laboral a lo más íntimo y personal, es un proceso diario tan normalizado que rara vez nos paramos a pensar en las implicaciones sociales —y en las preguntas morales— que la digitalización conlleva. Todos tenemos móviles inteligentes y, sin embargo, muy pocos conocemos sus tecnologías habilitadoras: el *big data* y la inteligencia artificial (en adelante IA). Confiamos en las tecnologías digitales a pesar de no conocerlas o entenderlas, y aun así experimentamos cada día las consecuencias de los cambios sociales y económicos que la IA basada en el big data está produciendo. Estos cambios plantean cuestiones morales cuyas consecuencias políticas y económicas generan un amplio debate sobre los principios éticos a seguir en el desarrollo de la IA.

Las controversias alrededor de la ética de la IA aparecen a diario. Por un lado, de las aplicaciones de IA dependen algunas de nuestras oportunidades en la sociedad: tanto de encontrar trabajo como de encontrar pareja. Nuestra reputación se mide por nuestra popularidad en las redes sociales. Nos preocupa la IA y el quedarnos obsoletos en un mundo diseñado por ingenieros informáticos y programadores, así como el efecto de

la desinformación sobre la capacidad de las democracias occidentales de perdurar y de no convertirse en regímenes autoritarios. También, la soberanía de nuestras naciones al haber infravalorado y no invertido en tener la capacidad de construir nuestros propios chips y microprocesadores o —en algún futuro no demasiado lejano— nuestros ordenadores cuánticos. Mirando más allá de la experiencia de las actuales generaciones que se formaron en la época analógica, como sociedad crece la preocupación por las consecuencias de la digitalización sobre las nuevas generaciones, es decir, sobre su bienestar y su capacidad de aprender, expresarse, amar y sentirse amadas.

La preocupación por la ética de la IA crece también entre científicos y profesionales que contribuyen activamente a la vanguardia de esta tecnología. Varios estudios ponen de manifiesto el papel de la IA en la automatización de la injusticia y en la perpetuación de dinámicas muy antiguas de discriminación y segregación social; nuevas formas de discriminación de género o de raza aparecen bajo la etiqueta de "sesgos algorítmicos". Las respuestas para este clima de preocupación incluyen códigos de conducta elaborados por comités de ética en las empresas u otros mecanismos de autorregulación. Crece la demanda de transparencia. Así, se pide abrir la caja negra de la IA para que aumente nuestra comprensión del uso que se les da a nuestros datos y de las consecuencias reales de las decisiones automatizadas. Crece también la demanda de regulación que limite, a través de mecanismo de competencias, el poder de las empresas tecnológicas al mando del oligopolio digital.

Este libro[1] presenta distintas controversias que han ido emergiendo, y cobrando cada vez más relevancia, alrededor

1. Este libro es parte de las actividades de divulgación del proyecto "Ingeniería de valores en sistemas de IA – VAE" (TED2021-131295B-C31-1) financiado por MCIN/AEI /10.13039/501100011033 y por la Unión Europea NextGenerationEU/PRTR.

de las implicaciones sociales y de los retos morales de la IA. Mientras que otros volúmenes[2] de esta colección hablan de aspectos específicos de estas tecnologías, este ensayo se centra en el impacto social y en los retos morales de la digitalización y de la economía basada en los datos.

La obra[3] permite reflexionar sobre esa confianza que tenemos en la IA y en las personas que están detrás de dicha tecnología. En el primer capítulo se definirá la IA y se contará brevemente su historia. En el segundo, hablaremos de los robots como expresión material de la IA. El tercer capítulo explica un tema central del debate sobre la IA: su fiabilidad y los riesgos y perjuicios que su uso puede generar para los individuos y la sociedad en su conjunto. Hablaremos de la relación entre bienestar y digitalización, intentando definir unas pautas básicas de comportamiento que nos pueden ayudar a proteger a nuestros seres queridos y a nosotros mismos de los riesgos de manipulación y engaño presentes en la red. En el cuarto capítulo hablaremos de los principios éticos que se están intentando adoptar en Europa para desarrollar una IA fiable. Finalmente, en el quinto capítulo se reflexiona sobre las soluciones que se están proponiendo tanto a nivel legal como técnico para promover el buen desarrollo de estas tecnologías y de las opciones que tenemos o que podríamos imaginar tanto como ciudadanos como profesionales y educadores, en cómo nuestras decisiones pueden influir en el desarrollo de la IA y del big data.

2. Véase, en esta misma colección, López de Mántaras y Meseguer (2017); Ríos y Gómez-Ullate (2019); Arroyo, Gayoso y Hernández (2020); Ríos y Naveiro (2022); García Armada (2022); Cobo y Lloret (2023).

3. Este libro no hubiera existido sin la paciencia y el cariño de Carmen Guerrero y Arantza Chivite. Quiero además agradecer sinceramente a Amalio Blanco, Jon Rueda y Daniel López Castro sus comentarios y sugerencias.

CAPÍTULO 1

¿Qué es la IA?

La inteligencia artificial representa un conjunto de ciencias (incluyendo la lógica matemática, la estadística, las probabilidades, la neurobiología computacional, la informática) que pretende imitar las capacidades cognitivas del ser humano. Este conjunto de teorías y técnicas se basa en la suposición de que todas las funciones cognitivas (el aprendizaje, el razonamiento, el cálculo, la percepción, la memorización e incluso el descubrimiento científico o la creatividad artística) pueden describirse con una precisión tal que sería posible programar un ordenador para reproducirlas.

Como vocablo y como disciplina se suele contar que la IA nació oficialmente durante un curso de verano organizado por cuatro investigadores estadounidenses[4] en 1956 en el Dartmouth College, en Hanover (Estados Unidos). Sin embargo, muchos historiadores consideran el libro *Cibernética* de Norbert Wiener (1948) el punto de partida de la disciplina (Nilsson, 2009) hasta que se produjo una división entre la cibernética y la IA sobre cuestiones relacionadas con los "sistemas simbólicos" y el papel de la psicología frente a la neurofisiología. Cada vez se trazaban más fronteras entre el

4. John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon.

modelado del cerebro y lo que se conoció como IA simbólica hasta el renacimiento de las redes neurales en la década de 1980 (Kline, 2010). Los investigadores al principio creían que la construcción de una inteligencia de nivel humano llevaría unos pocos años, un par de décadas como máximo. Ese optimismo inicial se desvaneció en la década de 1970 dando paso a los llamados "inviernos de IA", a los que siguieron nuevos momentos de esplendor. Podemos afirmar que desde hace más de medio siglo la IA sigue generando expectativas.

¿Piensan las máquinas?

Charles Babbage (1791-1871) trabajó hasta su muerte en el desarrollo de una "máquina analítica". En 1843, Ada Lovelace en sus "Notas" a la traducción de un artículo del ingeniero italiano Luigi Menabrea explicaba que la "máquina analítica" de Babbage, si se construía, sería un ordenador programable y no una simple calculadora. Tomaría datos de tarjetas perforadas y produciría resultados novedosos ejecutando diversas operaciones mecánicas paso a paso (Swetz, 2019)[5].

Ideas similares a las de Babbage se discutían también en España. En los *Ensayos sobre Automática* de Leonardo Torres y Quevedo (1914) se presenta el primer modelo de autómata que "ejecuta una por una todas las operaciones indicadas en la fórmula que se trata de calcular", que procede "en todo como un ser inteligente que sigue ciertas reglas", y, sobre todo, "en el momento en que hay que escoger un camino en cada caso particular" (González Redondo, 2019).

Por su parte, la patente estadounidense número 613 809 de Tesla describe el primer dispositivo de control remoto inalámbrico (Swezey, 1958). El modelo en funcionamiento o

5. Por su contribución a la obra de Babbage, Ada Lovelace ha sido considerada la primera mujer programadora de la historia.

"teleautómata" respondía a señales de radio y se alimentaba con una batería interna. En 1898, Nikola Tesla hizo una demostración de la primera embarcación de control por radio del mundo. Tesla no limitó su método a las embarcaciones, sino que generalizó el potencial del invento para incluir vehículos de cualquier tipo y mecanismos que se accionan con cualquier fin. Imaginó un operador o varios operadores dirigiendo simultáneamente 50 o 100 embarcaciones o máquinas a través de transmisores y receptores de radio sintonizados de forma diferente. Por desgracia, el invento estaba tan adelantado a su tiempo que quienes lo observaron no podían imaginar sus aplicaciones prácticas.

En 1942, John Vincent Atanasoff y su ayudante Clifford Berry crearon el ordenador Atanasoff-Berry (ABC) que pesaba unos 320 kg y podía resolver hasta 29 ecuaciones lineales simultáneas. En 1949, Edmund Berkeley —cofundador de la Association for Computing Machinery (ACM)— publica *Giant Brains: Or Machines That Think* (Cerebros gigantes o máquinas que piensan), en el que escribe: "Recientemente ha habido muchas noticias sobre extrañas máquinas gigantes que pueden manejar información con gran velocidad y habilidad […] Estas máquinas son similares a lo que sería un cerebro si estuviera hecho de *hardware* y cables en lugar de carne y nervios […] Una máquina puede manejar información; puede calcular, concluir y elegir; puede realizar operaciones razonables con información. Una máquina, por tanto, puede pensar".

¿De verdad podríamos decir que una máquina pueda pensar? ¿Basta con saber calcular para pensar? ¿Basta con pensar para tener conciencia? Y podríamos continuar con una larga disquisición sobre aquello que define al ser humano. Estas reflexiones nos llevan a pensar en el carácter fenoménico de la experiencia: es decir, en percepciones, en sensaciones corporales, en los estados de ánimo y las emociones sentidas.

El término *qualia* (Kind, 2021) fue introducido en la literatura filosófica (en su sentido contemporáneo) por Clarence Irving Lewis (1929) en una discusión sobre la teoría de los datos sensoriales para definir esa experiencia perceptiva que nos plantea el *duro problema de la conciencia* —también llamado laguna explicativa—, pues la imposibilidad de resolver el misterio de qué es lo que hay en el cerebro da cuenta de nuestra experiencia sensorial del mismo. Un problema, el de la conciencia, que ni la neurociencia ni la filosofía han sido aún capaces de resolver.

Sin embargo, según el exempleado de Google, Blake Lemoine, el generador de *chatbot* LaMDA (Language Model for Dialogue Applications, modelo de lenguaje para aplicaciones de diálogo) tiene conciencia propia —o, por lo menos, al interaccionar con él lo parece— (De Cosmo, 2022), y eso se debe a que el modelo responde a preguntas sobre temas que van desde la física hasta la filosofía. Los altos cargos de Google no comparten esas opiniones (Tiku, 2022) y por el momento nadie más en la empresa ha querido hablar de un tema que recuerda mucho a las películas de ciencia ficción.

Terminator, Robocop, Her, los replicantes de *Blade Runner* o Project 2501 en *Ghost in the Shell* son todas IA generales (o AGI) más o menos intangibles o robóticas. Gracias a las películas y a los libros de ciencia ficción estamos bastante familiarizados con AGI —la IA de una máquina que podría realizar con éxito cualquier tarea intelectual que pueda realizar un ser humano—. AGI se diferencia de ANI, es decir, de la IA específica o débil (*artificial narrow intelligence*). ANI es la IA programada para realizar una única tarea (ya sea comprobar el tiempo, ser capaz de jugar al ajedrez o analizar datos brutos para redactar artículos periodísticos). AGI en el mundo de la ficción respeta las tres leyes de la robótica de Isaac Asimov, diseñadas para evitar que los robots dañen a los humanos:

1. Un robot no hará daño a un ser humano o, por inacción, permitirá que un ser humano sufra daño.
2. Un robot debe cumplir las órdenes dadas por los seres humanos, a excepción de aquellas que entrasen en conflicto con la primera ley.
3. Un robot debe proteger su propia existencia en la medida en que esta protección no entre en conflicto con la primera o con la segunda ley.

Por su parte, ANI en el mundo de Meta Platforms (Facebook) en el que vivimos no tiene claro aún su código de conducta. Sin embargo, hoy en día muchos casos de éxito de aplicaciones de AI son ANI.

Entonces, ¿es LaMDA una AGI? Y si lo es, ¿tiene conciencia? Un experimento filosófico realizado por el académico John Searle en 1980 y llamado "la habitación china" nos ayuda a entender por qué creemos que Blake Lemoine se equivoca al decir que LaMDA es consciente. El experimento filosófico de Searle propone imaginar a un hombre sin conocimientos de chino dentro de una habitación al que se le introducen frases en chino por debajo de la puerta. El hombre manipula las frases de forma puramente simbólica (o, mejor dicho, sintáctica) según un manual de instrucciones que incluye un conjunto de reglas. Coloca respuestas que engañan a los de fuera haciéndoles creer que hay un hablante de chino dentro de la habitación. El experimento mental demuestra que la mera manipulación de símbolos no se puede considerar una forma de real comprensión. Sin embargo, la sofisticación a la que ha llegado el chatbot es tan elevada que podemos decir que pasaría cualquier test de Turing, es decir, que sería capaz de engañar a un humano haciéndose pasar por otro ser humano. Lo que lleva a la siguiente pregunta: ¿podría LaMDA actuar de forma intencional?

Autores como Nick Bostrom sostienen que los agentes artificiales inteligentes son capaces de razonar de forma

instrumental. Si asumimos, como hace él, que el razonamiento instrumental es sinónimo de inteligencia y afirmamos que la inteligencia es independiente de la motivación, podríamos afirmar como Bostrom que tanto AGI como ANI podrían perseguir objetivos propios, es decir, no compartidos por los seres humanos (como, por ejemplo, contar los granos de arena en una playa).

La IA podría perseguir varios valores instrumentales convergentes para una amplia gama de objetivos finales y una amplia gama de situaciones. Esto implicaría que estos valores instrumentales sean perseguidos por muchos agentes inteligentes para poder alcanzar un objetivo final de más alto nivel, que podría no tener sentido para una mente humana y, sin embargo, ser totalmente racional para una mente sintética. Esta situación generaría lo que, según el profesor de ciencia de la computación en la universidad de Berkeley Stuart Russell (2021), se denomina el "problema del control" o, mejor dicho, el problema de la falta de control del ser humano sobre las máquinas inteligentes.

Según Bostrom, si el desarrollo tecnológico continúa, en algún momento se alcanzará un conjunto de capacidades que harán que la devastación de la civilización sea extremadamente probable, a menos que la civilización no salga de la condición de semianarquía en la que vive por defecto. Sugiere también que para prevenir la devastación de la civilización se podría intentar impedir que se difunda la información peligrosa; restringir el acceso a los materiales, instrumentos e infraestructuras necesarios; disuadir a los posibles delincuentes aumentando las probabilidades de que los descubran; ser más precavidos y hacer más trabajo de evaluación de riesgos, o establecer algún tipo de mecanismo de vigilancia y aplicación que permita interceptar los intentos de llevar a cabo un acto destructivo (Bostrom, 2019).

¿Representa entonces la IA un riesgo existencial para la humanidad? ¿Son fundadas las preocupaciones de Bostrom?

Evidentemente, es difícil saber si un riesgo lo es hasta que no se materializa de alguna forma. De momento, las únicas muertes documentadas causadas por IA derivan de los accidentes de coches autónomos[6]. Podemos suponer la existencia en zonas de guerra de víctimas de sistemas de armas autónomas letales (LAWS, Lethal Autonomous Weapon Systems), también llamadas robots asesinos (*slaughterbots*). Varias organizaciones internacionales, entre ellas Cruz Roja, han intentado promover el debate[7] para restringir el uso de estos sistemas. Desafortunadamente, de momento no han tenido éxito principalmente por la oposición de Rusia y Estados Unidos a que se prohíba el desarrollo y uso de LAWS.

Evidentemente, podemos recurrir a nuestra capacidad de inferir lo que no sabemos a partir de lo que sí sabemos. Por ejemplo, que las sociedades humanas tienen una fuerte tendencia a resolver disputas a través de conflictos armados destructivos. Un análisis del catálogo de conflictos[8] elaborado por el doctor Peter Brecke del Georgia Institute of Technology indica que en los tiempos modernos las guerras son menos frecuentes, pero más destructivas (Martelloni, Di Patti y Bardi, 2018). Sin embargo, Steven Pinker (2011) sostiene que los cinco "demonios interiores" detrás de nuestra disposición a la violencia (depredación, dominación, venganza, sadismo e ideología) han disminuido a lo largo de los siglos a favor de cuatro "ángeles bondadosos" o capacidades que nos permiten "refrenar nuestros impulsos más oscuros" (empatía, autocontrol, pensamiento moral y razón), promoviendo el progreso humano.

A pesar de que el número de guerras ha disminuido a nivel mundial, siendo el siglo XVIII (el Siglo de las Luces) el periodo más evidente de paz y prosperidad, el aumento del poder

6. La muerte de Elaine Herzberg de 50 años fue el primer caso registrado de un peatón fallecido tras una colisión con un coche autónomo de Uber en Arizona, ocurrida el 18 de marzo de 2018.
7. Véase "Slaughterbots are here", en https://autonomousweapons.org/.
8. Véase https://brecke.inta.gatech.edu/research/conflict/.

destructivo de las guerras modernas es innegable: una guerra nuclear podría destruir la vida en la Tierra por completo. Entre las tecnologías contemporáneas, la IA juega un papel importante en los equilibrios de poder entre los Estados nacionales. La invasión rusa de Ucrania y la tensa relación entre Estados Unidos y China sobre Taiwán devuelven a la actualidad el debate y la necesidad de establecer mecanismos de cooperación diplomática y de control económico eficaces para prevenir la intensificación de los conflictos armados.

La IA entre guerras

Retomamos nuestro relato sobre la historia de la IA con Alan Turing[9] y el papel que jugó durante la Segunda Guerra Mundial un grupo de investigación liderado por él en Bletchley Park (Inglaterra). El principal objetivo del trabajo de Turing a partir de 1939 fue descifrar el código Enigma utilizado por las fuerzas armadas alemanas para enviar mensajes de forma encriptada. Turing y su grupo fueron capaces de descifrar Enigma gracias a la invención de una máquina electrónico-mecánica conocida como la "bomba" de Turing (figura 1).

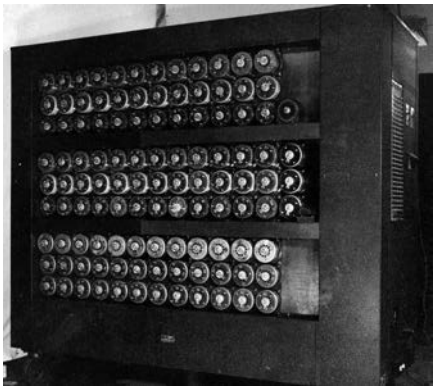
Los ordenadores digitales comparten el esquema básico de esta máquinaa[10]: un dispositivo de input/output o entrada/salida (cinta y lector), una unidad central de procesamiento (mecanismo de control) y una memoria (la unidad de almacenamiento del mecanismo de control) (Turing, 1937). Los ordenadores modernos son, en esencia, máquinas de Turing universales. A pesar de que Enigma había sido descifrado gracias a la colaboración de un equipo de ordenadores humanos e

9. La película *Descifrando Enigma* (2014) ofrece una reconstrucción de la vida de Turing.
10. En 1936, el matemático inglés Alan Turing propuso una herramienta matemática que podía reconocer aquellos enunciados matemáticos que, dentro de un determinado sistema formal de axiomas, no pueden demostrarse como verdaderos o falsos según los teoremas de incompletitud de Gödel.

informáticos, el antagonismo entre inteligencia artificial y humana iba a ser un motivo recurrente en el desarrollo de estas tecnologías. El término *inteligencia artificial*, acuñado por John McCarthy en 1955, contribuyó a aumentar las expectativas, fijando explícitamente el objetivo en una inteligencia artificial con capacidades similares a la humana.

Figura 1

Bomba de Turing usada para descifrar el código Enigma.



Fuente: Wikimedia Commons.

La rivalidad humanos-máquinas y el creciente interés en la resolución de problemas en la investigación en IA encontraron en el ajedrez el campo de batalla ideal. Ya en 1914, el ingeniero español Leonardo Torres y Quevedo había demostrado la viabilidad de construir la primera máquina de ajedrez, capaz de realizar finales de rey y torre contra rey sin intervención humana. En 1950, Claude Shannon[11] publicó el primer artículo sobre el desarrollo de un programa informático para jugar al ajedrez. En 1997, la IA ganó por fin la competición con los

11. Claude Elwood Shannon (1916-2001) era ingeniero de los laboratorios Bell en los que la informática ha sido una rama del Departamento de Matemáticas desde su creación en 1916.

humanos cuando el campeón mundial de ajedrez Garry Kasparov fue derrotado por Deep Blue de IBM. Y en 2016, el programa AlphaGo de Google se enfrentaba a Lee Se-dol en Seúl y ganaba el partido de Go[12] por 3 a 0.

Claude Shannon no ayudó solamente a que la IA supiera jugar al ajedrez, sino que fue el primero en reconocer que el álgebra de George Boole[13] podría utilizarse en los circuitos de relés de los conmutadores telefónicos para transmitir información abriendo el camino para su uso en todos los circuitos digitales. Gracias a su teoría matemática de la comunicación, la información se convertía en una unidad cuantificable que no necesita tener en cuenta el contenido del mensaje para ser transmitida. Los elementos básicos del modelo de Shannon (una fuente, un transmisor, un canal, un receptor, un destino y la posibilidad de que haya ruido) siguen siendo los mismos en todos los sistemas de comunicación, entre ellos internet.

En 1956, científicos geniales como Alan Turing, Norbert Wiener, Claude Shannon o Warren McCullough trabajaban de forma independiente en cibernética, matemáticas, algoritmos y teorías de redes. Fue el informático y científico cognitivo John McCarthy quien tuvo la idea de unir dichos esfuerzos de investigación en un solo campo: la inteligencia artificial. No solamente McCarthy acuñó el término, sino que también fundó laboratorios de IA en el MIT y en Stanford. Los historiadores de la IA han hablado a menudo de "ascenso y caída" para describir el auge del paradigma de la IA en los años cincuenta-sesenta y su aparente desaparición en las dos décadas siguientes (Agar, 2020). Sus *veranos* se caracterizaron por el optimismo y las inversiones, mientras que durante los *inviernos* se enfrentaron a recortes de financiación, incredulidad y pesimismo.

12. Go es un juego de dos jugadores que se turnan para colocar piedras negras o blancas en una cuadrícula de 19 por 19. Gana quien tenga el control de la mayor cantidad de territorio en el tablero, lo que se logra al rodear las piezas de su oponente con las suyas.

13. En 1854, Boole argumentó que el razonamiento lógico puede realizarse sistemáticamente de la misma manera que la resolución de un sistema de ecuaciones.

Los 17 años siguientes de la Conferencia de Dartmouth (1956) hubo increíbles avances gracias a proyectos de investigación llevados a cabo en el MIT, las universidades de Edimburgo, Stanford y Carnegie Mellon, y la financiación masiva que recibieron. La Agencia de Proyectos de Investigación Avanzada de Defensa de Estados Unidos (DARPA) invirtió millones de dólares en la investigación de la IA sin presionar a los investigadores para que obtuvieron resultados concretos. En 1951, Marvin Minsky y Dean Edmunds construyen el SNARC (Stochastic Neural Analog Reinforcement Calculator), la primera red neuronal artificial, utilizando 3000 tubos de vacío para simular una red de 40 neuronas. Las redes neuronales eran programas de IA inspirados en la estructura del cerebro, con muchas neuronas artificiales elementales con alta conectividad que implementan una función no lineal. La idea vino de un artículo escrito por Warren S. McCulloch y Walter Pitts en 1943, en el que los autores hablaban de redes de "neuronas" artificiales idealizadas y simplificadas, y de cómo estas podrían realizar funciones lógicas sencillas.

Los ordenadores empezaron a resolver problemas algebraicos, a demostrar teoremas geométricos y a utilizar la sintaxis y la gramática inglesas. Alentados por estos impresionantes primeros resultados, los investigadores estaban esperanzados en que para 1985 serían capaces de construir la primera máquina verdaderamente pensante, capaz de realizar cualquier trabajo que un hombre pudiera hacer. El problema era que los ordenadores no podían seguir técnicamente el ritmo de la complejidad de los problemas planteados por los investigadores; cuanto más complicado era el problema, mayor era la potencia de cálculo que requería. Por ejemplo, un sistema de IA que analizaba la lengua inglesa en aquel entonces solo podía manejar un vocabulario de 20 palabras, porque ese era el máximo de datos que podía almacenar la memoria del ordenador.

Los problemas técnicos y la falta de resultados visibles trajeron desilusión y recortes en la financiación. Por ejemplo,

la Enmienda Mansfield de 1969 exigía a DARPA que solo se financiara investigación directamente orientada a misiones concretas; es decir, que los investigadores solo recibirían financiación si sus resultados podían producir tecnología militar útil, como tanques autónomos o sistemas de gestión de batallas.

A principios de la década de 1970, la publicación del informe del Comité Asesor del Procesamiento Automático del Lenguaje (ALPAC), elaborado por el Gobierno de Estados Unidos en 1966, y el informe escrito por el profesor James Lighthill en Reino Unido y titulado *Artificial Intelligence: A General Survey*[14] sobre el desfase entre los resultados reales de la investigación en IA y las alocadas visiones de las máquinas pensantes dio lugar al "invierno de la IA", que dañó la credibilidad de los entusiastas de la IA y provocó una pérdida general de financiación. Cayó entonces el silencio sobre la IA hasta que, en los años ochenta, el interés se renovó, especialmente debido a la popularidad de los sistemas expertos.

Desde los sistemas expertos hasta el big data

El segundo *verano* de la IA coincide con la democratización de la informática mediante la difusión de los ordenadores de mesa. En 1984, Apple gastó un millón de dólares en un anuncio de 60 segundos que se estrenó en la Super Bowl. El vídeo, dirigido por Ridley Scott, terminaba con el famoso eslogan: "El 24 de enero, Apple Computer presentará Macintosh. Y verás por qué 1984 no será como *1984*"[15]. El Macintosh 512 KB, apodado Fat Mac, se presentó en septiembre de ese año y en solo cien días Apple vendió más de 72 000. El Fat Mac ofrecía a los usuarios cuatro veces más memoria y les permitía mantener abiertos simultáneamente varios programas. En 1987, Apple

14. Disponible en https://bitly.ws/VIEK.
15. *1984* es el título de la novela distópica escrita por George Orwell y publicada en 1949.

e IBM construían ordenadores de sobremesa mucho más potentes y baratos que las computadoras desarrolladas para trabajar en lenguaje de programación Lisp.

Mientras tanto, a nivel empresarial empezaban a difundirse los sistemas expertos, es decir, programas de IA que trataban de imitar el razonamiento de un experto humano en un ámbito concreto. En otras palabras, un sistema experto es un programa informático que, tras haber sido debidamente instruido por un profesional, es capaz de deducir información a partir de un conjunto de datos e información de partida. El programa informático Dendral, desarrollado en 1965, se considera el primer sistema experto porque automatizó el proceso de toma de decisiones y resolución de problemas de los químicos orgánicos en la Universidad de Stanford. En 1982, *el sistema experto* XCON, desarrollado por el profesor John McDermott para la Digital Equipment Corporation (DEC), permitía organizar los pedidos de ordenadores tomando en cuenta todas las combinaciones de opciones de configuración ofertadas por la empresa. Se dice que XCON hizo ganar a DEC unos 40 millones de dólares al año.

Desafortunadamente, la dificultad de redactar normas que reflejaran los conocimientos de los expertos y los problemas de mantenimiento de estos sistemas hicieron que los sistemas expertos llegaran a su fin. Además, en 1987 DARPA dejó de financiar la investigación en IA.

Empezaba entonces, a mediados de los años ochenta, otra época que durará hasta hoy: la de internet. En agosto de 1962, el doctor Joseph Carl Robnett Licklider, del MIT, expuso en sus notas su idea de una "red informática intergaláctica": un conjunto de ordenadores interconectados a escala mundial a través del cual todo el mundo podría acceder rápidamente a datos y programas desde cualquier lugar. Como director de la Agencia de Proyectos de Investigación Avanzada (ARPA) del Departamento de Defensa de EE UU, Licklider fue uno de los promotores en 1969 de ARPAnet —la primera red informática

pública de conmutación de paquetes—. En 1972, la red ya estaba operativa con 40 puntos conectados en diferentes localizaciones (Leiner *et al.*, 1997). En octubre de 1980, ARPAnet sufrió una interrupción de servicio de cuatro horas que se considera el primer ejemplo de denegación de servicios (DDoS en inglés). Cuando en 1982 adoptó el protocolo TCP/IP, se creó internet (International Net). Fue desmantelado en 1990.

Desde internet y el big data hasta el más allá

En menos de 30 años, Internet ha cambiado la forma de vivir de muchas personas, llegando a la mayoría de los hogares en los países desarrollados. En diciembre de 1995, internet tenía 16 millones de usuarios; en diciembre de 2022, 5544 millones. En España, en 2010, solo la mitad de los hogares tenía internet; 12 años después, en 2022, la banda ancha llegaba casi a la totalidad de las casas. Según el INE[16], en 2020 en España el 95% de los hogares tenían conexión de banda ancha y subía un 6,9% respecto a 2019 el número de personas que compraban *online* (53,8%), año en el que el 87% de las mujeres españolas entre 16 y 74 años declaraba usar el móvil para navegar por internet (EUROSTAT, 2021).

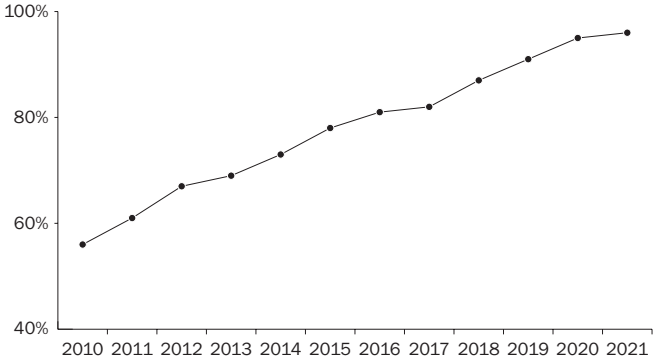
La rápida adopción de internet durante los años noventa llevó a que estallara la burbuja dotcom a finales de esa década. De aquellas cenizas se levantaron los que hoy conocemos como los gigantes tecnológicos: Alphabet (Google), Amazon, Facebook y Apple (GAFA). Google se fundó el 4 de septiembre de 1998 por los informáticos Larry Page y Sergey Brin cuando eran estudiantes de doctorado en la Universidad de Stanford. El nombre Google proviene de *gúgol*, es decir, 10 elevado a 100 o 1 seguido de 100 ceros. En mayo de 2023, Alphabet[17] tenía una

16. "Equipamiento y uso de TIC en los hogares - Año 2020", en https://bitly.ws/VIRB.
17. Google representa más del 99% de los ingresos anuales totales de Alphabet.

cotización de 1340 miles de millones de dólares, ocupando el cuarto puesto detrás de Saudi Aramco, Microsoft y Apple, y delante de Amazon, NVIDIA, Meta y Tesla.

Figura 2

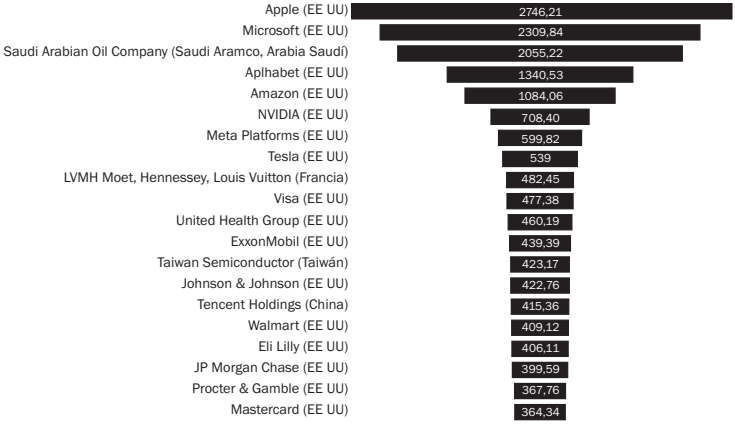
Hogares con acceso a internet en España entre 2010 y 2021.



Fuente: Elaboración propia a partir de datos de Eurostat.

Figura 3

Mayores empresas del mundo por valor de mercado en 2023.



Fuente: Elaboración propia a partir de datos de *Forbes*, 5 de mayo de 2023, en Statista.

El mundo de la IA que conocemos hoy se basa en una simple constatación: la importancia de los datos, concretamente el big data, para el éxito de la IA. En el mundo empresarial, el origen del término big data se atribuye a un artículo publicado en el año 2001 por Doug Laney, vicepresidente de investigación de Gartner Research, en el que destacó tres tendencias del mundo empresarial: 1) el notable volumen de datos transaccionales generados por el comercio electrónico y la voluntad de las empresas de conservar esta información; 2) la velocidad de creación de datos producida por la interacción entre organizaciones y clientes; y 3) la oportunidad de integrar y gestionar una mayor variedad de información, con formatos y estructuras diferentes. Estos elementos constituyen las famosas 3 V del big data: su volumen, variedad y velocidad de generación de datos por parte de usuarios finales, así como de entidades públicas y privadas.

Según el catedrático de teoría de las organizaciones Thomas Davenport (2014), el big data viene a indicar una nueva forma de competición de mercado basada en el análisis automatizado de grandes conjuntos de datos, lo que permite a aquellos que manejen esa tecnología mejorar la eficiencia de sus procesos, personalizar sus servicios y anticipar futuras demandas de sus clientes. El big data, en otras palabras, representa una fuente de ventaja competitiva y un nuevo estadio de lo que anteriormente se definía como *business intelligence* y que hoy llamamos *big data analytics*, es decir, la intersección entre las tecnologías de almacenamiento y gestión de datos con el mundo del aprendizaje automático y de las redes neuronales. El big data representa entonces la intersección de varias tecnologías que permiten vincular y comparar grandes conjuntos de datos y utilizar esos datos para identificar patrones y tomar decisiones.

Si en 2010 el big data era el tema de conversación más común en las reuniones de negocios, en el año 2023 el valor económico y social del big data y de la IA específica son incuestionables. Al ser un concepto muy vago, el big data coincide

también con la creencia generalizada de que los grandes conjuntos de datos digitales ofrecen una forma superior de inteligencia que puede generar un conocimiento más objetivo, neutral y preciso (Boyd y Crawford, 2012). Una "mentalidad basada en big data" también parece justificar que sea aceptable que las redes sociales al mismo tiempo midan, manipulen y moneticen el comportamiento de los usuarios. Dentro de la economía de la atención basada en el refrán "si no pagas por el producto, tú eres el producto" parece normal que los usuarios faciliten información personal a las empresas y reciban servicios a cambio.

El aprendizaje automático es la parte de la IA que está influyendo más directamente en nuestras vidas y que podemos comparar con otros inventos o descubrimientos determinantes como la electricidad o el fuego. Al ser el aprendizaje automático excelente para reconocer patrones en los datos, se puede usar en cualquier ámbito de aplicación. Así, los ornitólogos lo utilizan para clasificar el canto de los pájaros, los astrónomos para buscar planetas en los destellos de la luz de las estrellas y los bancos para evaluar el riesgo de crédito y prevenir el fraude. El estado de avance de áreas como el procesamiento del lenguaje natural (NLP) o el procesamiento de imágenes es tan alto que sus resultados —desde la escritura automática de artículos a los *deepfakes*— nos parecen alarmantes.

El capitalismo de la vigilancia

Nuestros datos constituyen el activo más importante de la economía digital, también llamada economía de las plataformas o capitalismo de la vigilancia (Zuboff, 2019). Acontecimientos como los confinamientos asociados a la pandemia de COVID-19 han fortalecido y amplificado la expansión del ecosistema digital de datos. Según la Comisión Nacional de los Mercados y la Competencia (CNMC), en el segundo trimestre de 2020, la facturación de los sitios de comercio electrónico aumentó

en España 12 020 millones de euros. El ecosistema digital en el que vivimos depende de las decisiones de multinacionales y gigantes tecnológicos que definen cada día más nuestra forma actual de vivir siguiendo las dinámicas de un mercado fundamentalmente oligopolístico en el que 11 empresas estadounidenses tienen hoy en día más de 100 millones de abonados digitales. En el año 2000, pocos predijeron el carácter monopolístico de la tecnología digital; hoy en día, hay amplio consenso entre los economistas sobre este tema.

La Unión Europea ha marcado entre sus prioridades la autonomía y soberanía tecnológica en su estrategia digital para 2025. Si miramos a las empresas tecnológicas europeas, vemos que SAP sigue siendo el titán tecnológico más poderoso por ingresos, con unas ventas anuales de casi 30 000 millones de euros. Su valor de mercado es de 123 000 millones de euros y ha celebrado su 50º aniversario en abril de 2022. La longevidad de SAP se debe en parte a que una vez que las empresas optan por un determinado tipo de *software* empresarial resulta complicado sustituirlo. Sin embargo, es uno de los pocos gigantes que ha sobrevivido a tres cambios radicales: de los ordenadores centrales a los sistemas "cliente-servidor" más distribuidos, luego a Internet y ahora a la computación en la nube (*cloud computing*). Entre las razones por la que SAP es de los pocos gigantes europeos en el ámbito digital se mencionan la reacia actitud al riesgo de empresarios y consumidores europeos, la falta de capital de riesgo, la burocracia y un mercado interno fragmentado. Todos problemas conocidos y que siguen sin solución.

La economía de las plataformas

El debate sobre los retos sociales de la IA incluye autores que critican la economía del dato y que hablan de *datificación* pidiendo justicia (de los datos) (Dencik *et al.*, 2019); también los que hablan de economía de las plataformas (Mansell y

Steinmueller, 2020) —o de la vigilancia— proponiendo soluciones para romper o regular los oligopolios. Los economistas consideran que, al ser intermediarias, las plataformas digitales tienen la tendencia a convertirse en oligopolios. A esto se suma la facilidad de adoptar una estrategia de innovación basada en la adquisición de empresas más pequeñas. ¿De quién estamos hablando? De los cinco gigantes digitales estadounidenses cuyos productos y servicios "alegran" nuestras vidas: Apple y Microsoft, que tienen más de 40 años, Alphabet (Google) y Amazon, que han superado los 20 años, y Facebook, que pasó a llamarse Meta a partir de octubre de 2021 y que cumplió 18 años en febrero de 2022. Estos gigantes tenían en 2021 un valor de mercado de 7,6 billones de dólares; se estima que sus ventas se duplicarán en la próxima década. Para que nos hagamos una idea de lo que significa, pensemos que el presupuesto de gasto del Estado español para 2022 fue de 347 486 millones de euros.

Es difícil competir en el mercado de las plataformas digitales: hay demasiados lazos estrechos entre empresas. Alphabet paga a Apple hasta 12 000 millones de dólares al año para que Google sea el motor de búsqueda por defecto del iPhone. Mirando más allá de Estados Unidos, vemos que China cuenta con Alibaba y Tencent y otros cinco contendientes con un valor de 100 000 millones de dólares o más. Huawei representa una alternativa al duopolio Apple-Alphabet de sistemas operativos iOS-Android. La India cuenta con Jio y el sudeste asiático con Grab, Gojek y Sea. Alibaba y Tencent tienen participaciones en algunas de las nuevas empresas chinas. China tiene una nueva lista de "las nueve cosas que no hay que hacer" para las empresas de comercio electrónico, entre las que se incluye no dejar fuera a los nuevos competidores. Se piensa que el resurgimiento de las autoridades antimonopolio puede marcar la diferencia en el futuro de la economía digital. Para eso, Europa está planeando normas para que los productos de las distintas empresas sean interoperables y permitan a los usuarios mover sus datos entre las distintas plataformas.

Evidentemente, el panorama del mercado digital es cambiante. La proporción de ingresos de los cinco gigantes americanos que se solapan con los demás ha pasado del 22 al 38% desde 2015[18]. Por ejemplo, Microsoft y Alphabet se enfrentan a Amazon en la nube (*cloud*). Amazon es, a su vez, la fuerza en ascenso en publicidad digital. Además, la cuota de mercado del resto de empresas se mantiene estable, en torno al 35%, en cada uno de los 11 subsectores tecnológicos estadounidenses. Pero la cuota de las segundas y terceras empresas ha aumentado del 18 al 26% desde 2015: se reduce la distancia entre los gigantes y sus competidores. PayPal, por ejemplo, aspira a tener 750 millones de usuarios de su aplicación financiera para 2025. Empresas tecnológicas independientes, como Shopify en el comercio electrónico, se han abierto paso y están generando suficientes beneficios para ser autosuficientes. Nuevas versiones de Steve Jobs emergen cada día reorientando el rumbo de lo que llamamos innovación.

Para terminar

Este capítulo nos ha servido como punto de partida para situar nuestra conversación sobre los retos morales y sociales de la IA en el contexto de la economía de las plataformas, también conocido como capitalismo de la vigilancia. La digitalización ha generado la materia prima que alimenta el desarrollo de los modelos de IA basados en aprendizaje automático. Estos modelos, que tanto éxito están teniendo, inundan los dispositivos digitales que colonizan cada día más cada detalle de nuestras vidas. En este breve ensayo intentaremos encontrar un hilo narrativo que nos permita relacionar debates y aplicaciones muy diversas: desde las redes sociales y la IA generativa a los robots y sus usos duales.

18. "The rules of the tech game are changing: A new phase in the global tech contest is under way", *The Economist*, 27 de febrero de 2021.

CAPÍTULO 2

¿Puede ser ética la IA?

Cuando hablamos de ética de la IA queremos que la ética aplicada (o práctica) nos ayude a saber cómo deberíamos actuar en determinadas situaciones. La proliferación de nuevas aplicaciones de IA y la rapidez en la adopción de las mismas hace que sea difícil predecir *a priori* los problemas que generará la IA en cada caso. Esta situación suele hacer que las reflexiones surjan cuando los problemas se convierten en escándalos que preocupan a la opinión pública. Titulares de prensa del tipo "Cómo los algoritmos te discriminan por origen racial y por género" promueven este debate y hacen que queramos saber qué hacer en el caso de una IA que, a pesar de cumplir con el objetivo para el que ha sido desarrollada, reproduzca sesgos incrustados generando resultados discriminatorios. Evidentemente, las preocupaciones (más o menos fundadas) que genera la IA son directamente proporcionales a su nivel de penetración en nuestras vidas. Por eso, en los siguientes capítulos hablaremos de artefactos digitales muy diversos, mientras que dedicaremos este capítulo a hablar de los principios éticos y filosóficos que nos pueden ayudar a enfrentarnos a los retos que pone la IA.

Una definición preliminar de ética de la IA

La ética consiste fundamentalmente en formular juicios morales. La metaética es la forma más abstracta de la ética y trata sobre si existen o no verdades morales. La ética descriptiva intenta describir lo que un grupo concreto de personas considera correcto o incorrecto, sin relacionarlo necesariamente con ninguna teoría subyacente o concepción global de la moralidad. La ética normativa, por su parte, se centra en cómo debe actuar la gente y es el ámbito de la ética en el que se debaten y definen las tres principales teorías éticas (utilitarismo, deontología y ética de la virtud). Por último, está la ética aplicada, que son las teorías éticas normativas aplicadas a circunstancias particulares (por ejemplo, la ética médica, la empresarial, la de la investigación o la asistencial). Según Sætra y Danaher (2022), la ética aplicada incluye también distintas formas de ética de la tecnología, y en esta podemos incluir la ética de la ingeniería, la ética de la tecnología, la ética de la informática, la ética de la IA o la ética de la robótica.

La ética de la IA abarca cuestiones como el diseño y uso de sistemas autónomos en general (tanto armas como otros sistemas), los prejuicios de las máquinas, la privacidad y la vigilancia, la gobernanza, el estatus de las máquinas inteligentes, la automatización y el desempleo, e incluso la colonización espacial. No hay además consenso sobre la validez de denominar esta área de estudio como ética de la IA. Floridi (1999) sugería hablar de ética de la información, Langford (2000) de ética de internet, y Anderson y Aderson (2011) de ética de las máquinas. Recientemente, se ha propuesto hablar de ética de los datos (Hand, 2018) y de ética digital (Luke, 2018), reconociendo la contribución de los estudios sobre medios de comunicación digitales, los estudios de vigilancia y privacidad y la sociología crítica feminista, entre otros.

En su manifestación más cercana a la filosofía, la ética de la IA se expresa a través de principios y directrices que

puedan guiar el diseño, el despliegue o la adopción de la IA. Al ser un área nueva y en evolución, esas directrices van evolucionando cada día. Dicha evolución refleja los hallazgos presentados en revistas y eventos académicos[19], así como los cambios normativos presentes en códigos de conducta o en leyes. También hay una parte de ética normativa que la profesora Casey Fiesler, en su análisis de los temarios de los cursos de ética de la tecnología impartidos en escuelas de ingeniería informática, resume en el siguiente objetivo formativo: la habilidad de detectar problemas y criticar lo establecido considerando alternativas de diseño, o de implementación, que combinen múltiples perspectivas epistémicas (Fiesler, Garrett y Beard, 2020).

La capacidad de realizar un análisis crítico del diseño, implementación e impacto de la IA es lo que previene el sesgo de automatización, es decir, la tendencia de las personas a usar de forma sesgada las sugerencias proporcionadas por un sistema automatizado de recomendación (Skitka, Mosier y Burdick, 1999). En este sentido, el diseño de un sistema de IA puede aumentar o disminuir el riesgo de que se produzca el sesgo de automatización. Un estudio sobre sistema de co-razonamiento humano-máquina muestra que los 204 participantes del mismo eran más propensos a seguir la lógica del sistema de IA cuando se les daba una respuesta prefabricada, pero cuando la IA planteaba una pregunta, los participantes admitían que el sistema de IA les hacía cuestionarse más sus reacciones y les ayudaba a ser más críticos y reflexivos (Danry *et al.*, 2023).

Hoy en día, la difusión de los sistemas de recomendación en todos los sectores nos da la impresión de que son las máquinas las que tienen el poder de tomar decisiones cotidianas que alteran nuestras vidas. Cuando se permite que sea la IA la

19. Como, por ejemplo, el congreso "FAccT – Justicia, responsabilidad y transparencia" organizado por la Association for Computing Machinery (ACM), en Estados Unidos.

que determine en segundos la solvencia de los solicitantes de préstamos o se ofrece a los jueces recomendaciones sobre el riesgo que alguien que ha cometido un delito en el pasado vuelva a delinquir, nos enfrentamos a nuevos posibles episodios de sesgos de automatización. Vista la gravedad de estos asuntos, el hecho de que muchos de estos sistemas de IA sean "cajas negras", es decir, que los seres humanos no pueden acceder ni comprender fácilmente la lógica que lleva al sistema a recomendar una determinada decisión, hace difícil poder sondear y cuestionar lo que el sistema de IA propone, además de que se fomenta cierta complacencia y pleitesía hacia la máquina.

Al ser la IA un campo muy complejo, hacen falta intermediarios que ayuden a los usuarios finales a entender las limitaciones y a usar de forma correcta estos sistemas. El rechazo por parte de las empresas privadas de poner sus algoritmos de IA a disposición de los usuarios para que lo examinen, porque consideran que el software es propiedad intelectual, es otro tipo de falta de transparencia que exige que se establezcan procedimientos de auditorías de algoritmos con el fin de reducir el riesgo de que se produzcan resultados adversos.

Otro problema bastante conocido es el de los sesgos de los algoritmos, es decir, que los datos que utilizamos para "entrenar" algoritmos reflejan y perpetúan las injusticias ya presentes en la sociedad. El famoso estudio de Joy Buolamwini y Timnit Gebru (2018) —retratado en el documental de Netflix de 2020 *Sesgo codificado*— identifica una falta de precisión en los modelos de reconocimiento facial asociado con la baja presencia de información de determinados grupos sociales en las bases de datos. El estudio evidencia una mayor precisión del algoritmo a la hora de identificar rostros de hombres con tono de piel más claro en comparación con mujeres con tono de piel más oscuro. La sospecha de discriminación abre el camino a un análisis tecnopolítico de la IA.

Un estudio de 113 conjuntos de datos para procesamiento de imágenes evidencia cuatro valores que guían los autores que construyen estas bases de datos: eficacia, universalidad, imparcialidad y calidad de los modelos de aprendizaje automático (Scheuerman, Hanna y Denton, 2021). Para cada valor, los autores identifican un valor silenciado, es decir, aquellos que se devalúan implícitamente en favor de los valores adoptados. Estos valores silenciados son: la paciente curación de los datos, su contexto, su posición social y política y la calidad de las prácticas adoptadas en generar la base de datos.

Este tipo de estudios critican la neutralidad u objetividad de los algoritmos poniendo de relieve los valores epistémicos y políticos de que están embebidos los algoritmos. El estudio de los sesgos algorítmicos permite apreciar la grave disparidad racial y de género de muchos sistemas basados en el aprendizaje automático y saber, por ejemplo, que los sistemas de reconocimiento del habla malinterpretan a los hablantes negros el doble de veces que a los blancos (Koenecke *et al.*, 2020).

Cualquier sistema automático de recomendación puede tener este tipo de sesgos más allá de los sistemas de reconocimiento de imágenes o de voz. Se trata, por ejemplo, de que un algoritmo que examina solicitudes de empleo pueda establecer sus criterios para determinar qué tipo de currículum es el adecuado basándose en quién tiene esos puestos en ese momento. Si en la base de datos con la que se entrena el algoritmo hay una mayoría de hombres blancos, es probable que el algoritmo decida que el candidato ideal es un hombre blanco. Una vez que tengamos este conocimiento, hay que asegurarse de que los profesionales involucrados no se abstengan de actuar creyendo que son otros los responsables de intervenir. Además, no debemos creer que el sesgo algorítmico depende solo de deficiencias del conjunto de datos: a menudo, los problemas pueden atribuirse a la dificultad de medir las interacciones entre todas las variables usadas en un modelo (Hooker, 2021).

La ética de la IA en Europa

Durante los últimos años, muchas entidades públicas y privadas se han esforzado en elaborar líneas maestras para el desarrollo ético de la IA. Muchas de estas iniciativas añaden nuevos principios a los cuatro clásicos de la bioética —una de las áreas más avanzadas de la ética aplicada—, que son: beneficencia, no maleficencia, autonomía y justicia. Es decir, que la IA debe no solamente contribuir a la promoción del bienestar individual y del bien común general, sino también superar los prejuicios potenciales que pueda generar su uso para los sujetos, las comunidades y el medioambiente. Por eso, es preciso añadir al principio de *no maleficencia* los principios de *precaución* y *prevención*, intentando asegurar que la IA no haga daño a nadie o a nada. Además, la IA ha de estar centrada en las personas y respetar el derecho de las mismas a decidir y elegir libremente. Podemos también interpretar el concepto de *justicia* tanto como *equidad* como *no discriminación*. La puesta en práctica de todos estos principios requiere establecer mecanismos de control que permitan reparar los daños generados de forma voluntaria o involuntaria por parte de los actores responsables del despliegue del sistema de IA. Podemos entender estos principios como un amparo y escudo frente a la fragilidad y vulnerabilidad que podemos experimentar como individuos y como sociedad con relación a la IA (Ausín, 2021).

La reflexión ética representa entonces una invitación a analizar los valores y las creencias de todos los agentes (por ejemplo, diseñadores, programadores, desarrolladores, etc.) involucrados en el diseño y despliegue de estas tecnologías produciendo una metaética, es decir, un análisis del origen de los principios éticos que aplicaremos durante el proceso de coproducción de nuestro sistema de IA y así impedir que los grupos involucrados (programadores, usuarios, clientes, autoridades de control, etc.) caigan en el error de pensar que el problema no es de ellos, sino de otros.

En Europa, la Lista de Evaluación de la Inteligencia Artificial Confiable (ALTAI) pretende ayudar a evaluar si el sistema de IA que se está desarrollando, desplegando, adquiriendo o utilizando cumple con los siete requisitos detallados en 2022 en las "directrices éticas para una IA de confianza" elaboradas por el Grupo de Expertos de Alto Nivel en Inteligencia Artificial de la Comisión Europea:

1. El respecto de la autonomía de los humanos y su papel como supervisores.
2. La robustez y seguridad técnica de los sistemas de IA.
3. El respecto de la privacidad de los datos mediante mecanismos de gobernanza adecuados.
4. La transparencia e inteligibilidad de la IA.
5. El respecto de la diversidad y de la equidad en contra de toda forma de discriminación.
6. La protección del bienestar ambiental y social.
7. La rendición de cuentas.

ALTAI en la práctica representa una lista de comprobación, es decir, un recordatorio de aspectos a tener en cuenta a la hora de evaluar la confiabilidad de un sistema de IA en desarrollo. El éxito de su implementación depende del nivel de preparación y compromiso de los agentes involucrados en el diseño, el despliegue y el uso del sistema de IA. Volviendo a lo anterior: el lugar en el que se manifiesta y resuelve el dilema moral vuelve a ser la cadena de producción del algoritmo, software o servicio de IA. Miremos ahora algo más en detalle cada uno de estos principios.

ALTAI #1: la protección de la autonomía del ser humano

El primer principio pide que la IA no sustituya a las personas, sino que les permita tomar decisiones informadas y gozar de todos sus derechos fundamentales. Al tal fin, deben existir

mecanismos de supervisión basados en enfoques como por ejemplo el de *human-in-the-loop*. En otras palabras, aunque sean ayudados por una IA, los seres humanos deben conservar la autonomía de debatir las sugerencias de la IA y decidir finalmente por ellos mismos.

El artículo 22 del Reglamento General de Protección de Datos (RGPD) dice explícitamente que "todo interesado tendrá derecho a no ser objeto de una decisión basada únicamente en el tratamiento automatizado, incluida la elaboración de perfiles, que produzca efectos jurídicos en él o le afecte significativamente de modo similar", a menos que dicha decisión no sea parte de un contrato entre las partes, esté autorizado por ley o el interesado haya dado su consentimiento explícito a dicha operación. Este artículo no solamente impone un principio de transparencia y visibilización del uso de métodos automatizados de recomendación, sino también la facultad de que los interesados puedan recurrir la decisión con la ayuda e intermediación de un operador humano. Este principio pretende salvaguardar la libertad y autonomía del ser humano, contrarrestando la pleitesía y subordinación a la máquina.

ALTAI #2: garantías de robustez y seguridad técnica de la IA

Los sistemas de IA deben ser resistentes y seguros, garantizando un plan de contingencia que permita la continuidad de las operaciones en caso de que surjan problemas inesperados. Que la IA sea robusta implica que sus resultados sean exactos, fiables y potencialmente reproducibles. Asegurar la calidad de los resultados y prevenir cualquier intento de manipulación de los mismos con, por ejemplo, ataques de aprendizaje automático adversario (*adversarial machine learning*), es fundamental para prevenir que el sistema vulnere los derechos o los intereses del usuario final en línea con el principio de no maleficencia heredado de la bioética. Asegurarse de que los

sistemas de IA sean seguros implica evitar su uso indebido, su manipulación y velar sobre su correcto funcionamiento. En otras palabras, la IA no debe dañar a nadie o a nada.

ALTAI #3: protección de la privacidad y buen gobierno de los datos

Vista la importancia del aprendizaje automático, es evidente el papel central que los datos y su gestión juegan en el éxito de cualquier proyecto de IA. Como muchos de estos datos pueden definirse como datos personales, su protección (que en EE UU se llama privacidad de la información) se convierte en una obligación en la Unión Europea al ser un derecho fundamental establecido en el Tratado de Lisboa.

El RGPD de la Unión protege el derecho fundamental a la protección de los datos personales estableciendo el marco jurídico sobre la protección y la libre circulación de los datos personales dentro de la UE. Cualquier entidad pública o privada que se dedique al tratamiento de datos personales de ciudadanos europeos o cuyo establecimiento principal esté en la UE debe cumplir con el RGPD.

En España, la Ley Orgánica de Protección de Datos y Garantía de los Derechos Digitales (LOPDGDD) complementa el RGPD, matizando algunos aspectos específicos de la ley dentro del marco jurídico español. Estas leyes imponen que los responsables y encargados del tratamiento adopten mecanismos adecuados de gobernanza de los datos, teniendo en cuenta su calidad e integridad y garantizando el acceso legítimo a los mismos.

El buen gobierno de los datos implica además preservar la relación de confianza entre los que ceden y los que procesan los datos, tema especialmente importante a la hora de establecer, por ejemplo, espacios de datos sanitarios[20].

20. Véase la guía de la AEPD de mayo de 2023, "Aproximación a los espacios de datos desde la perspectiva del RGPD", en https://bitly.ws/VMQm.

Evidentemente, el derecho a la intimidad en el entorno virtual va más allá de un buen modelo de gobernanza de datos. Si, por ejemplo, creemos como Cory Doctorow[21] que "la privacidad es el derecho a equivocarse", entendemos fácilmente por qué el artículo 94 del RGPD prevé el derecho de supresión —conocido como derecho al olvido— que limita la difusión y con ella la memoria perpetua —en internet o en las redes sociales— de ciertos acontecimientos de los que fuimos protagonistas y que nos gustaría olvidar.

ALTAI #4: promoción de la transparencia

Los sistemas de IA deben diseñarse de tal forma que no sean cajas negras, sino que se mantenga cierta inteligibilidad y conocimiento por las partes interesadas de su lógica interna y de las fuentes y modelos de datos y de negocio que influyen en su funcionamiento. Los usuarios y operadores deben ser conscientes de que están interactuando con un sistema de IA y deben conocer las capacidades y limitaciones del mismo.

ALTAI #5: obligación de no discriminar asegurando la equidad y respetando la diversidad

Los sistemas de IA no deben reproducir prejuicios sociales que puedan marginar a grupos vulnerables o desembocar en la discriminación de sus miembros. Para fomentar la diversidad, los sistemas de IA deben ser accesibles y todos y todas deben poder participar en su diseño y desarrollo.

ALTAI #6: promoción del bienestar ambiental y social

Los sistemas de IA deben beneficiar a todos los seres humanos, incluidas las generaciones futuras. Por ello, hay que

21. En https://bitly.ws/Wdr6.

garantizar que sean sostenibles y respetuosos con el medioambiente.

ALTAI #7: obligación de establecer mecanismos de rendición de cuentas

Deben establecerse mecanismos que permitan identificar responsabilidades y sanciones en caso de que se genere algún perjuicio o daño como consecuencia del uso de un sistema de IA. Es además necesario auditar la IA, esto es, evaluar los algoritmos, los datos y los procesos de diseño de la IA para asegurarse de que no haya sesgos incrustados. Además, deben establecerse mecanismos de compensación adecuados que permitan que haya una reparación de los daños.

Limitaciones de una ética de la IA basada en principios

Una revisión de la literatura de los principios éticos para la IA en 2020 identifica siete áreas temáticas recurrentes y en general coherentes con ALTAI: 1) control humano de la tecnología, 2) responsabilidad y promoción de valores, 3) privacidad, 4) seguridad, 5) transparencia y explicabilidad, 6) equidad y no discriminación, 7) compromiso profesional (Fjeld *et al.*, 2020). A pesar del consenso alrededor de estos principios, algunos expertos critican la validez de adoptar un enfoque basado en principios y motivan su escepticismo diciendo que la IA carece de: objetivos comunes y deberes fiduciarios, códigos profesionales, métodos probados para llevar los principios a la práctica, y sólidos mecanismos jurídicos y profesionales de asunción de responsabilidad (Mittelstadt, 2019). Además, parece superfluo celebrar el consenso en torno a principios de alto nivel que siguen ocultando profundos desacuerdos

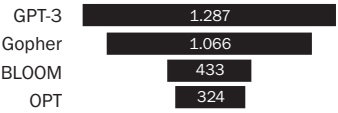
políticos y normativos en cada contexto específico de aplicación.

Por eso, mientras que los enfoques éticos occidentales canónicos, desde la deontología hasta el consecuencialismo, proponen teorías y principios para la IA universales y generalizables, teorías como el afrofeminismo (Collins, 2002) hablan de "interseccionalidad" entre sexismo, opresión de clase y racismo, reivindicando la necesidad de analizar la IA poniendo el foco en la estrecha relación entre prácticas socioculturales y desarrollo tecnológico (Birhane *et al.*, 2022). Romper estas prácticas requiere un cambio social profundo que no puede materializarse simplemente en un código de buenas prácticas. Sirva como ejemplo un estudio (Akbar *et al.*, 2023) en el que se muestra que el 84% de las investigadoras e investigadores que han publicado en las actas del congreso de ACM FAccT sobre equidad, transparencia y responsabilidad entre 2018 y 2022 replican dinámicas de poder al provenir de países occidentales y de contextos cultos, industrializados, ricos y democráticos. Hasta el Foro Económico Mundial reconoce que los beneficios económicos y sociales de la IA siguen concentrados geográficamente, sobre todo en el Norte global. En el Sur global, la falta de disponibilidad y compatibilidad de los datos y una conectividad de red fiable siguen siendo importantes puntos débiles.

En su libro, Kate Crawford (2021) —investigadora de Microsoft Research y catedrática de IA y Justicia en la École Normale Supérieure de París— revela el entramado económico que permite el desarrollo de las tecnologías detrás de la IA. Crawford empieza su cartografía de la IA por la famosa mina de metales de tierras raras de Mountain Pass en Norteamérica. La IA se presenta como una tecnología extractiva cuya vida depende de la energía eléctrica (figura 4) y cuya creación necesita muchos minerales y metales, pero también el esfuerzo físico y mental de miles de seres humanos.

Figura 4

Consumo de energía para el entrenamiento de LLM en 2023 (en megavatios/hora).



Fuente: Elaboración propia a partir de datos de Luccioni *et al.* (2022), en Statista.

"Hacer solo el bien" (beneficencia) y "no hacer daño" (no maleficencia), aunque parezcan a primera vista dos formas de hablar de lo mismo, representan en realidad ideas fundamentalmente distintas. La IA puede contribuir al bienestar de ciertos grupos o a enriquecer a unos pocos poderosos en detrimento de comunidades que ya han sido desposeídas en el pasado. La IA, como cualquier tecnología que transforma recursos naturales, estará siempre en el ojo del huracán de los conflictos de interés entre grupos sociales. Un debate centrado únicamente en principios abstractos no permite entender la materialidad y con ella la complejidad, los conflictos y las inercias asociadas a la producción de la IA. Sin dicha comprensión es imposible proponer medidas adecuadas para prevenir o resolver los problemas generados por la creciente adopción de tecnologías digitales, que, además, plantea preguntas y problemas éticos (Heilinger, 2022) del tipo: ¿sobre quién recae la responsabilidad en caso de daños derivados del uso de un sistema de IA? ¿Cómo puede evitarse que los sistemas de IA reflejen la discriminación, los prejuicios y las injusticias sociales existentes a partir de sus datos de entrenamiento, agudizándolos de esta manera? ¿Cómo podemos comprender, explicar y controlar —si es que alguna vez podemos hacerlo— el funcionamiento interno de un sistema complejo de IA? ¿Cómo puede protegerse la intimidad de las personas, dado que los datos personales pueden ser recogidos

y analizados con tanta facilidad? ¿Cómo puede protegerse la autonomía de las decisiones humanas y contrarrestar la influencia indebida y la pérdida de capacidades humanas (*deskilling*) resultantes del uso de la IA?

Para responder a estas cuestiones hay autores que proponen la formulación de principios que orienten sobre lo que moralmente debe hacerse, con base en catálogos de responsabilidades y deberes; otros proponen sopesar las promesas frente a las consecuencias perjudiciales para centrarse en un uso "virtuoso" de la IA o incluso en una IA virtuosa en sí misma. Finalmente, autoras como yo misma sostienen que el desarrollo ético de la IA es facultad y responsabilidad de todos los actores involucrados (desde los mismos diseñadores y desarrolladores de estas tecnologías hasta sus usuarios).

La primera línea de la ética de la IA

En el diseño y desarrollo de la IA quien seguramente tiene estatus ontológico y moral son sus diseñadores y desarrolladores y con ellos y ellas también quienes contribuyen con sus datos y su uso a su perfeccionamiento. Es decir, que quienes se enfrentan cada día a los dilemas morales que genera la IA somos nosotros. Por eso, la ética de la IA incluye tanto cuestiones deontológicas como de filosofía de la tecnociencia porque, como contamos a continuación, la ética de la IA es también un problema de ética de la investigación.

Así, en 2012, el Departamento de Seguridad Nacional de EE UU publicó el "Informe Menlo: Principios éticos que guían la investigación en tecnologías de la información y la comunicación", donde se pide que antes de empezar cualquier proyecto TIC es necesario aclarar la relación entre investigadores, sujetos y tecnología. Hay que respetar la autonomía de los individuos que participan en el desarrollo tecnológico, tutelando a las personas más vulnerables.

Los investigadores deben tener en cuenta que los riesgos para los sujetos se sopesan contra los beneficios para la sociedad, y no en beneficio de los investigadores individuales o de los propios sujetos de investigación. Los investigadores deben actuar con la debida cautela teniendo en cuenta las obligaciones establecidas en leyes, reglamentos, contratos y otros acuerdos privados relevantes para su investigación. Los comités de ética de la investigación deben analizar cada caso con atención, especialmente aquellos en los que no se pueda obtener consentimiento informado por parte de las personas que participan directamente, indirectamente o con sus datos en la investigación o en el desarrollo de la IA.

Consecuentemente, la ética de la IA también incluye la deontología profesional del programador. En el Código de Ética y Conducta Profesional de la ACM de EE UU se declara que "un profesional de la informática debe: 1.1. Contribuir a la sociedad y al bienestar humano, reconociendo que todas las personas son partes interesadas en la informática. 1.2. Evitar el daño, es decir, cualquier consecuencia negativa, especialmente cuando esas consecuencias son significativas e injustas. 1.3. Ser honesto y digno de confianza, es decir, proporcionar información sobre las capacidades, pero también las limitaciones y problemas potenciales de un sistema […] 1.4. Ser justos y no discriminar, es decir, fomentar la participación equitativa de todas las personas, incluidas las de los grupos infrarrepresentados […]".

Finalmente, la ética de la IA trata temas de responsabilidad social corporativa que a su vez reflejan la deontología de las empresas públicas y privadas protagonistas de su despliegue. Este es probablemente el contexto en el que el disenso y la desobediencia salgan más caros. Tomando como ejemplo el respeto del principio de seguridad y robustez técnica en el desarrollo de la IA, podemos pensar que las empresas digitales tienen todo el interés por cumplir estos principios. Sin

embargo, la evidencia empírica demuestra que a veces los intereses de negocio generan perversos desincentivos para invertir en ciberseguridad.

Este es el caso de Peiter Zatko, jefe de seguridad de Twitter desde noviembre de 2020 hasta enero de 2022, y cuyo trabajo como *hacker* ético se conoce bajo el pseudónimo "Mudge". En 2022, a los 51 años, Zatko decidió poner en riesgo su brillante carrera profesional para revelar secretos sobre las dudosas actividades llevadas a cabo en Twitter con el objetivo de producir un cambio importante en la empresa. Según Menn (2022), en la demanda enviada a la Comisión del Mercado de Valores en julio de 2022, Zatko habla de Twitter como de una empresa paralizada por un liderazgo deshonesto y sin rumbo, una empresa contaminada por influencias extranjeras, acosada por "atroces" fallos de privacidad y seguridad, en fin, un peligro para la seguridad nacional y susceptible incluso de un colapso total.

En la demanda oficial se puede leer lo siguiente sobre las motivaciones de su denuncia: "Cuando los hackers éticos encuentran una vulnerabilidad que los malos pueden explotar, primero proceden a hacer una 'divulgación responsable' silenciosa para que la empresa o el gobierno afectados puedan arreglarla. Pero a veces la organización vulnerable no quiere escuchar la verdad ni solucionar el problema. En esos casos, los hackers éticos se ven obligados a sopesar los riesgos de una divulgación más amplia: exponer las vulnerabilidades pone en alerta a personas malintencionadas, pero también permite a los usuarios tomar decisiones informadas y así permitir una mejora del servicio".

Hay tantos casos de exempleados y exempleadas denunciando malas prácticas, especialmente en el contexto de la economía digital, que desde junio de 2023 es obligatorio también en España establecer en las empresas y organismos públicos canales anónimos de denuncia. Esta obligación viene de la Ley 2/2023 reguladora de la protección de las personas

que informen sobre infracciones normativas y de lucha contra la corrupción, la que transpone la Directiva europea 1937/2019 para la protección de las personas que informen sobre infracciones del derecho de la Unión.

Para terminar

Hoy en día, el debate sobre la ética de la IA incluye tanto cuestiones de "justicia de datos", sesgo algorítmico y activismo digital como temas de diseño de sistemas de IA que estén alineados con los valores de los humanos que los usen y preguntas filosóficas fundamentales sobre conceptos como los de autonomía y responsabilidad. En las páginas de este capítulo hemos intentado tratar varios temas relacionados con este debate con la intención de fomentar la curiosidad más que para dar respuestas concluyentes.