Artificial

La nueva inteligencia y el contorno de lo humano

Mariano Sigman
Santiago Bilinkis

DEBATE

Sigman, Mariano
Artificial / Mariano Sigman ; Santiago Bilinkis - 1ª ed.
- Ciudad Autónoma de Buenos Aires : Debate, 2023.
232 p. ; 23 x 15 cm. (Debate)

ISBN 978-987-795-066-3

1. Divulgación - Ciencias aplicadas. I. Bilinkis, Santiago.
II. Título.
CDD 540

Mariano Sigman:
A mis viejos y a Michita, con amor y gratitud

Santiago Bilinkis:
A la memoria de mis abuelos: Anita, Bernardo, Rosita y Raúl, que aún viven en mi recuerdo; y por el futuro de mis nietos, que algún día llegarán

Primera edición en la Argentina bajo este sello: octubre de 2023

Printed in Argentina – Impreso en la Argentina

ISBN: 978-987-795-066-3

Queda hecho el depósito que previene la ley 11.723.

Compuesto en M. I. Maquetación, S.L.

Esta edición de 20.500 ejemplares se terminó de imprimir en Gráfica Pinter S.A., Diógenes Taborda 48, Ciudad de Buenos Aires, en el mes de septiembre de 2023.

Índice

Prólogo

Mariano Sigman

Un día de primavera en Madrid, Emiliano Chamorro me sugirió agregarle un capítulo a un libro que yo había escrito sobre la conversación, uno que hablase de cómo conversar con una inteligencia artificial. Ahí empezó todo. La idea viajó a la velocidad de un rayo hasta Miguel Aguilar y Roberto Montes, editores, maestros y amigos de uno y otro lado del Atlántico. Y volvió, casi al tiempo que Emi terminaba su frase, con otra propuesta: «¿Por qué no mejor un libro nuevo?». Y en esa sucesión vertiginosa se resolvió que el libro, además, tendría que escribirse en un instante.

Sin pensarlo un segundo, levanté el teléfono, llamé a Santiago, con quien siempre nos merodeamos pero con quien nunca había colaborado, y le propuse escribir, a cuatro manos y con una fecha bastante inminente, un libro. Era como invitar a alguien con el que nunca se ha salido a caminar a subir juntos el Everest. Como nada era normal en ese día, resultó que Santiago justo acababa de empezar no sé cuántos proyectos que ya en sí mismos parecían un abismo, y mientras yo ahí iba pensando que ese rapto de locura duraba lo que duran esos raptos, dijo que sí, que no sabía cómo, pero que ahí fuéramos. Y ahí fuimos.

Santiago Bilinkis

La llamada de Mariano me encontró en un momento de desborde total: al trabajo habitual como divulgador en la radio y a la generación de contenido para mi podcast y las redes, se sumaba el repentino interés de los medios por entender la revolución del ChatGPT. La inteligencia artificial, un tema al que le dedico gran parte del tiempo desde los últimos quince años y que solo nos interesaba a unos pocos «nerds», estaba de repente en el centro de la agenda pública. La gran meta de esta etapa de mi vida, acercar la tecnología más avanzada a la vida de las personas de una manera que les resulte sencilla y estimulante, cobraba más relevancia que nunca. Lo último que necesitaba en ese momento era una propuesta tan extraordinaria como indeclinable. Y justo me sonó el teléfono. Mi respuesta a Mariano fue instantánea. La idea de trabajar juntos me resultaba muy estimulante y combinar nuestras ideas en un proyecto conjunto era una oportunidad maravillosa. Pero no se acababa ahí: pocas cosas me atraen tanto como una meta imposible. Hacer un libro sobre inteligencia artificial, escribiendo por primera vez de a dos, con un coautor que vive en otro país con cinco horas de diferencia horaria, y completarlo en unas pocas semanas... ¡Imposible! ¿Dónde hay que firmar?

1

La génesis de la inteligencia

Tragedia y esperanza

En mayo de 1938, el almirante Sir Hugh Sinclair del Servicio de Inteligencia Británico, el mítico MI6, compró una mansión construida en el siglo XIX conocida como Bletchley Park. Este lugar era el emplazamiento ideal para crear un centro de operaciones: estaba a poco más de setenta kilómetros de Londres, cerca de una línea de tren que pasaba por las universidades de Oxford y Cambridge, y el esplendor arquitectónico del palacio ayudaría a camuflar las actividades secretas del gobierno durante la Segunda Guerra Mundial.

Poco después, el Servicio de Inteligencia fue «de pesca» a las universidades más importantes del Reino Unido para reclutar a un formidable equipo de treinta y cinco físicos y matemáticos, que serían liderados por Alan Turing y Dillwyn Knox. Así, de manera abrupta y precipitada, se puso en marcha esta sucursal secreta de la Escuela de Códigos y Cifrado del Gobierno del Reino Unido. Apenas llegaron a los espléndidos jardines de Bletchley Park, el grupo de nerds descubrió cuál sería su misión: ni más ni menos que salvar al mundo. Los alemanes utilizaban una máquina llamada «Enigma», que encriptaba sus mensajes a través de un sofisticado sistema de engranajes basado en tres rotores que transformaban cada letra en otra. El objetivo de Turing y su equipo era descifrar ese código. La tarea era extremadamente difícil ya que los nazis cambiaban a diario la posición inicial de los rotores y eso resultaba en 159 trillones de combinaciones posibles. Había que volver a

descifrar la posición cada vez. Desencriptar estos mensajes podía inclinar la balanza de la Segunda Guerra Mundial, porque permitiría a los aliados acceder a información reservada sobre los planes y acciones enemigos.

Como gran parte de los hombres jóvenes estaban destinados al campo de batalla, el gobierno británico reclutó a más de seis mil mujeres para trabajar en Bletchley Park. Hablaban varias lenguas, y eran muy hábiles jugando al ajedrez y resolviendo crucigramas. Entre ellas estaba Joan Clarke, que se convirtió rápidamente en una de las personas decisivas del proyecto.

Turing, Clarke y su equipo trabajaban contrarreloj, urgidos por el avance del conflicto bélico. Pasadas algunas semanas, descubrieron cómo descifrar los mensajes. Pero tan pronto entendieron los cálculos y decisiones necesarios para descifrar el código de Enigma, descubrieron también que era imposible que pudieran resolverlos a tiempo. Encontraron la solución en otro de los recintos de Bletchley Park, donde el propio Turing estaba desarrollando una máquina de cálculo que recibió el nombre de «Bombe». Con la ayuda de este enorme dispositivo electromecánico, creado en 1939 a partir de un viejo proyecto del matemático polaco Marian Rejewski, sería posible determinar el contenido de los mensajes encriptados por la máquina Enigma.

Los códigos nazis se descifraron a tiempo gracias a una asombrosa conjunción de factores humanos y tecnológicos: por un lado, un equipo privilegiado de mentes científicas que pasaron, sin previo aviso, de explorar universos abstractos en una pizarra a salvar el mundo, y de personas que convirtieron su afición por los enigmas y crucigramas en el principal recurso para descifrar el contenido de los mensajes secretos del Reich. Por otro, parece ser que la insistencia vanidosa de los nazis en usar repetidamente la fórmula «Heil Hitler» fue un error garrafal que simplificó la tarea, ya que es mucho más sencillo descifrar un código en el que hay mensajes previsibles que se repiten. Y por último, la puesta a punto de dispositivos aparatosos capaces de ejecutar a gran velocidad cálculos que los cerebros combinados de esos científicos no hubiesen realizado a tiempo.

Bombe no hubiese pasado una prueba de inteligencia. Ejecutaba apenas un cálculo demandante y sofisticado para descifrar un enigma. Pero este esbozo de pensamiento humano depositado en un dispositivo eléctrico mostraba ya algunos rasgos de lo que identificamos como inteligencia. Podía hacer operaciones y tomar decisiones que hasta ese momento solo realizaban personas «inteligentes». El programa que ideó Turing para establecer la posición inicial de los rotores de Enigma fue una versión muy rudimentaria de una inteligencia artificial (IA).

Así, en estos días, en los que suele percibirse la IA como algo opuesto a lo humano, quizá convenga recordar que su primer proyecto embrionario se concibió justamente en la urgencia por salvar a la humanidad de su poder de autodestrucción.

¿Qué Dios detrás de Dios la trama empieza?

Bletchley Park dejó de funcionar en cuanto terminó la guerra. Los matemáticos y físicos que habían sido reclutados en las mejores universidades volvieron a casa como los soldados que regresan del servicio. Pero, a diferencia de estos últimos, los héroes y heroínas de Bletchley Park no pudieron decir dónde habían estado ni qué habían hecho, y cargaron de por vida con el peso de ese secreto. Así, un manto triste y oscuro cubrió el final de esta historia épica.

Como si un descubrimiento clave para el desenlace de la Segunda Guerra Mundial no hubiese sido suficiente para una vida, Turing continuó sus investigaciones en los temas más intrincados y desafiantes de la ciencia. En un trabajo seminal titulado «La base química de la morfogénesis», publicado siete años después de terminar la guerra, reveló el mecanismo que da lugar a los patrones sofisticados de la naturaleza, desde la forma de las flores, o las de una célula, hasta las espirales de los caracoles. Siguiendo esta premisa, la de mostrar que las cosas más sorprendentes de la vida emergen de reglas sencillas, reanudó la tarea de entender la inteligencia. Y se propuso emularla, retomando el proyecto que había empezado en Bletchley Park.

Pasada la urgencia bélica, Turing entendió que el ajedrez, un juego que históricamente ha funcionado como una metáfora del ingenio humano, era un terreno idóneo para estudiar la inteligencia en un dominio acotado pero significativo. «Dios mueve al jugador, y éste, la pieza /¿Qué Dios detrás de Dios la trama empieza / de polvo y tiempo y sueño y agonías?», escribe Borges en un poema que repara en la misma analogía. El ajedrez se convirtió, desde ese momento, en el conejo de indias de la historia de la IA, fue el primer gran escenario para su exploración y desarrollo, y en la actualidad es el mejor terreno para observar qué sucede cuando una inteligencia sobrehumana se asienta en alguno de nuestros dominios.

«¿Cómo se diseña un programa capaz de analizar una posición de ajedrez y con criterio para tomar buenas decisiones?», se preguntó Turing. Para entender cómo funcionan los mecanismos de la inteligencia se basó en sí mismo. Analizó sus razonamientos para intentar comprenderlos y extrapolarlos a una máquina. Este fue el primer paso en el camino de búsqueda de una IA: emular y replicar la inteligencia humana. Más precisamente, la inteligencia de Turing. Este ejercicio de pensar sobre nuestro propio pensamiento, conocido como «metacognición», hasta el momento solo había interesado a la psicología, como una búsqueda de hacer explícito el proceso mediante el cual razonamos.

Turochamp, el primer programa de ajedrez, nació en 1948 a partir de una investigación que Turing y David Champernowne compartieron en Mánchester. El programa funcionaba como una receta de cocina. Una serie de instrucciones secuenciadas definían los pasos para decidir un movimiento de manera tan bien especificada que podría usarlo cualquier persona, aunque nunca hubiera jugado al ajedrez.

A Turing le ocurrió algo similar a lo que había vivido Leonardo Da Vinci en el siglo XV: su genio estaba adelantado al desarrollo tecnológico de la época. Como Turochamp estaba por encima de las capacidades de *hardware* disponible, no contaba con computadoras capaces de ejecutar el programa que había diseñado. Advirtió entonces que una forma de resolver el problema era ejecutar el programa en su cerebro, llevando adelante una tras otra las instrucciones que le indicaba el algoritmo. Turochamp fue el primer programa de IA y se ejecutó en un cerebro humano.

Su habilidad para jugar resultó bastante mediocre. Además, como programa, tenía grandes limitaciones: operaba sobre un único dominio específico con reglas muy claras (jugaba al ajedrez, pero no podía hacer ninguna otra cosa, ni siquiera jugar a un juego mucho más sencillo), y dependía de la claridad del lenguaje formal usado por el programador, de su imaginación y su conocimiento del juego.

El proyecto, además, nació con una herida de muerte: conocer en detalle cómo operaba el programa y cómo resolvía cada decisión lo volvió menos atractivo, ya que gran parte de la fascinación que nos produce la inteligencia humana radica, precisamente, en no entenderla. Algo que no es enigmático, sorprendente e inexplicable no nos parece inteligente. Y, por su misma estructura, el programa que Turing había diseñado carecía de estos elementos.

Turochamp fue un hito histórico pero sus resultados nunca fueron muy prometedores. Sesenta y cuatro años después, en 2012, en el marco de la celebración del centenario del nacimiento de Turing, la Universidad de Mánchester rescató el algoritmo que él había creado y lo enfrentó a uno de los mejores jugadores de todos los tiempos: Garry Kaspárov. El gran maestro ruso aplastó al viejo programa en una partida de dieciséis movimientos.

La frontera de lo humano

En 1950, Turing publicó un artículo académico en el que presentó, por primera vez, el andamiaje teórico de la prueba que conocemos como «test de Turing». ¿Pueden pensar las máquinas?, o para hacer la pregunta algo más precisa: ¿pueden pensar de una forma indistinguible a como lo hace un ser humano?

Turing propuso un método para resolver este interrogante, basado en el juego de imitación. En esta prueba, un entrevistador alterna preguntas a través de un terminal a dos *entes*: uno es una persona y otro, una computadora. Si el interrogador consigue dis-

tinguir quién es la persona y quién la máquina, el sistema no pasa la prueba de Turing. Si, por el contrario, la computadora logra confundirlo, entonces la supera.

El test de Turing es ingenioso y establece un criterio conciso para evaluar la inteligencia de las máquinas, pero tiene varios problemas. Se basa en una idea antropomórfica, ya que asume que una inteligencia artificial tiene que asemejarse a una humana. Además, ser capaz de camuflarse no es condición necesaria ni suficiente para ser inteligente. Aunque seamos más inteligentes que un chimpancé no podríamos hacernos pasar por uno, y por ende no hubiésemos pasado su test de Turing. De la misma forma, puede haber IA muy potentes que no logren emular la inteligencia humana, y otras que resemblen la inteligencia humana sin que por eso sean inteligentes.

En una trágica ironía, a Turing, que salvó al mundo de un horror inenarrable y dedicó su vida a estudiar el razonamiento, lo condenó la irracionalidad moral de su época. Turing era homosexual, lo que por aquel entonces se veía como una desviación de la salud mental y un peligro para la sociedad. En 1952, robaron en su casa y, en el marco de la investigación, lo presionaron hasta hacerle confesar que el ladrón era un amante. La víctima del robo pasó a ser un acusado de «indecencia grave», y en lugar de cumplir la pena en la cárcel —lo que le habría costado su puesto de investigador— se adhirió a una condena con libertad condicional y fue sometido a un proceso de castración química que consistió en una serie de inyecciones de estrógenos para reducir su libido sexual.

El 7 de junio de 1954, encontraron el cuerpo sin vida de Turing. Dicen que a su lado había una manzana a medio comer a la que, según se sospecha, le había inyectado previamente cianuro. Con solo cuarenta y dos años, una de las mentes más brillantes de la historia de la humanidad, un verdadero promotor del mundo libre y democrático, murió acorralado por ese mismo mundo libre que aún estaba repleto de prejuicios. En la génesis de la historia de la IA hay una profunda tragedia humana.

El equilibrio nuclear

Durante la Segunda Guerra Mundial, la ciencia invadió el terreno de la política a un lado y otro del Atlántico. Mientras Turing, Clarke y una tropa de mujeres lingüistas, matemáticas e incluso expertas en crucigramas impulsaron la computación, la criptografía y la IA en Bletchley Park, del otro lado del océano Albert Einstein escribía su célebre carta a Franklin D. Roosevelt, entonces presidente de Estados Unidos, que comenzaba así:

> Algunos trabajos recientes de E. Fermi y L. Szilard, que me han sido comunicados en forma manuscrita, me llevan a suponer que el elemento uranio puede convertirse en una nueva e importante fuente de energía en el futuro inmediato... Este nuevo fenómeno conduciría también a la construcción de bombas, y es concebible, aunque mucho menos seguro, que puedan construirse así bombas extremadamente poderosas de un tipo nuevo.

Para explorar el potencial bélico de este tipo de armas se diseñó el Proyecto Manhattan cuyo centro de operaciones era el Laboratorio Nacional de Los Álamos, en Estados Unidos, una instalación que, de la misma manera que Bletchley Park en Inglaterra, funcionaba en secreto y congregaba a los mejores cerebros de la época. Allí, los físicos más destacados en mecánica cuántica y física atómica, liderados por Robert Oppenheimer, trabajaban en el desarrollo de una bomba nuclear. Estos científicos también jugaron un factor decisivo en el resultado de la guerra.

Con el fin del conflicto y el triunfo de los aliados, las tecnologías de estos dos proyectos tomaron caminos muy dispares. La IA se convirtió durante varias décadas en un campo de estudio periférico, curioso pero intrascendente, que interesaba solo a un grupo minoritario de entusiastas de la tecnología y de la ciencia ficción. Por el contrario, el armamento nuclear se convirtió en el eje fundamental para el balance geopolítico, y en el factor decisivo de la Guerra Fría.

Estados Unidos y la Unión Soviética desarrollaron sendos pla-

nes de armas nucleares y ambas potencias llegaron a contar con la misma capacidad de reducir la civilización a cenizas. Esta paridad sirvió, en varias ocasiones, para frenar una escalada bélica de potenciales consecuencias terribles. Era un equilibrio nefasto y lleno de incertidumbre, sin duda, pero un equilibrio al fin. El matemático John von Neumann, que también fue uno de los grandes pioneros de la computación y la teoría de juegos, describió matemáticamente este equilibrio con una fórmula que interpela a la razón: $1 + 1 = 0$.

Lo llamativo es que a esto no se llegó de manera azarosa. En la década de 1940, muchos de los expertos en física nuclear de Estados Unidos compartieron sus saberes con la otra gran potencia, en un fabuloso *thriller* de espionaje. Algunos científicos se convirtieron en espías de la Unión Soviética porque apoyaban los ideales comunistas, pero otros lo hicieron sobre la base del concepto de «paridad nuclear». Vislumbraron el futuro y entendieron que, para evitar la aniquilación del planeta, había que asegurarse de que ningún país tuviera el monopolio de ese poder destructivo. Así, el precario equilibrio entre las dos potencias enfrentadas fue el resultado de la decisión de un grupo muy pequeño de científicos que entendieron que el conocimiento debía estar en manos de ambos para así lograr esa situación de tablas: la doctrina conocida como «destrucción mutua asegurada». El estadounidense Ted Hall, licenciado en Harvard, y el científico inglés Klaus Fuchs, fueron exponentes cabales de esta teoría de la paridad nuclear. Convencidos de que había que igualar las condiciones de juego para salvar a la humanidad de un desastre, contactaron con los soviéticos y los mantuvieron informados sobre los avances del Proyecto Manhattan.

La visión de este grupo de científicos, que entendieron que la distribución de tecnología nuclear iba a determinar el futuro del mundo, y que ellos tenían un rol decisivo e inevitable —por acción u omisión— en la configuración del mapa global, puede servir como guía para pensar acerca del avance de la investigación y el control sobre la IA en el futuro cercano. Veremos en este libro que hoy el poder de influencia de los dos proyectos tecnológicos ha cambiado: ya no será el poder destructivo de la energía nuclear sino

el de la IA el que ocupe el foco principal de la escena política y económica.

Eliza, el primer vestigio humano en una máquina de silicio

Mientras el mundo estaba pendiente de la tensión áspera entre dos potencias nucleares, la IA seguía ocupando un lugar bastante marginal en la esfera de las preocupaciones sociales. Por aquel entonces, la IA no era ni de cerca un ámbito de ricos y famosos. En sus reductos académicos, físicos, matemáticos y neurofisiólogos como Marvin Minsky, John Hopfield o Warren McCulloch comenzaron a trabajar en la idea de «redes neuronales», un concepto que permitió vislumbrar cómo emerge la inteligencia a partir de un sustrato que no es inteligente.

En este nuevo abordaje ya no se observaba la inteligencia humana para escribirla en un programa, sino que se pretendía ver si un cerebro digital y artificial era capaz de producir comportamientos inteligentes. Así, se superaba una limitación fundamental del planteamiento de Turing, ya que la mayoría de las cosas que hacemos involucran mecanismos que son inaccesibles para nosotros mismos. La búsqueda de la inteligencia a través de redes neuronales no dejaba de ser una concepción antropocéntrica, pero implicaba un cambio rotundo.

Lo asombroso del cerebro humano no radica en la complejidad de una neurona, sino en las capas y formas en las que se organizan miles de millones de ellas. Una neurona tiene, en esencia, una tarea muy simple: escucha a otras, y si éstas producen una señal suficientemente fuerte, entonces dispara y envía esa señal a otras vecinas. Así se forma un circuito entre unidades muy simples capaz de codificar una gran cantidad de patrones en las distintas configuraciones de neuronas encendidas y apagadas. Pronto veremos cómo las redes neuronales se convirtieron en el motor de la IA, pero antes nos toca presentar a una de las primeras celebridades de esta disciplina. Se llama Eliza. Y es un programa.

En los mismos años en que Minsky y Hopfield daban cuerda a una incipiente ciencia de la inteligencia, el psicólogo norteamericano Carl Rogers andaba preocupado por otra dimensión del pensamiento humano: la locura. Con su colega Abraham Maslow, cambiaron la filosofía y la práctica de la psicoterapia, dándole una perspectiva más humanista. En una época en la que la locura era vista como una enfermedad en la que las personas perdían aquello que los hacía humanos, Rogers propuso un acercamiento empático en psicoterapia y advirtió que la locura nos atraviesa a todos. La locura, poco a poco, empezó a verse, y a tratarse, como una de las tantas facetas y expresiones en el mundo diverso y variopinto de lo humano. Rogers hizo de la empatía el centro de la relación terapéutica.

Estos dos universos, el de la psicoterapia y la empatía por un lado, y el de las redes neuronales y la IA por el otro, pueden parecer muy distantes, pero en realidad son bastante cercanos. Ambos se ocupaban de entender, por construcción u observación, la esencia de lo humano. No fue casual que estos dos mundos se encontraran en uno de los primeros hitos de la IA. En 1966, inspirado por la noción de empatía que había desarrollado Rogers, medio siglo antes del ChatGPT, Joseph Weizenbaum, un profesor de informática del Instituto Tecnológico de Massachusetts (MIT), programó Eliza, el primer bot conversacional de la historia. Un intento de sintetizar a un psicoterapeuta rogeriano, intrínsecamente empático.

Eliza conversaba por escrito con su interlocutor e intentaba dar la impresión de que comprendía sus problemas e inquietudes. Para lograr que la conversación fuese empática y lo más humana posible, el programa se valía de un algoritmo simple. Primero, reconocía alguna palabra clave y de esa palabra podía inferir el tema. Por ejemplo, cuando el interlocutor mencionaba la palabra «padre» o «madre», Eliza reconocía que estaba hablando sobre la familia. Luego, hacía alguna pregunta o comentario relacionados con ese tema. Además, tenía almacenadas frases hechas, y también usaba frases de continuidad, que eran útiles para simular que mostraba interés.

A continuación, un ejemplo de cómo era un diálogo con Eliza:

MUJER JOVEN: Los hombres son todos iguales.
ELIZA: ¿En qué sentido?
MUJER JOVEN: Siempre nos están molestando por algo en concreto.
ELIZA: ¿Se te ocurre algún ejemplo concreto?
MUJER JOVEN: Bueno, mi novio me hizo venir aquí.
ELIZA: ¿Tu novio te hizo venir aquí?
MUJER JOVEN: Dice que estoy deprimida la mayor parte del tiempo.
ELIZA: Lamento oír que estás deprimida.
MUJER JOVEN: Es verdad. Soy infeliz.
ELIZA: ¿Creés que venir aquí te ayudará a no ser infeliz?
MUJER JOVEN: Necesito ayuda, eso seguro.
ELIZA: ¿Qué significaría para vos recibir ayuda?
MUJER JOVEN: Quizá podría aprender a llevarme bien con mi madre.
ELIZA: Hablame más de tu familia.
MUJER JOVEN: Mi madre me cuida.
ELIZA: ¿Quién más de tu familia te cuida?
MUJER JOVEN: Mi padre.
ELIZA: Tu padre.
MUJER JOVEN: Te pareces a mi padre en algunas cosas.
ELIZA: ¿Qué parecido ves?
MUJER JOVEN: No eres muy agresivo, pero creo que no quieres que me dé cuenta de eso.

Este programa, instalado en una computadora tan gigantesca como primitiva, basado en unas pocas líneas de código de una sencillez aplastante, resultó ser una estrella de la conversación. Todos querían hablar con Eliza. Más allá de su destreza circense y de ser la prueba de que era posible que una máquina de silicio conversase, demostraba que la empatía, y con ella uno de los rasgos esenciales de la condición humana, es mucho más simple de lo que creemos. Un programa rudimentario, que simplemente propone a una persona continuar hablando sobre el mismo tema, genera la ilusión de ser empático.

Pero Eliza, como Turochamp, no pasaría una prueba de inteligencia rigurosa. Era incapaz de memorizar, no podía aprender de sus conversaciones, no entendía la ironía, había un sinfín de temas sobre los que no podía opinar, y su concepción sobre qué era comprender a su interlocutor radicaba simplemente en continuar proponiendo una conversación sobre un mismo tema. Tampoco pasaría el test de Turing, pero sí podría engañar a su interlocutor durante un rato simulando algo profundamente humano. Y había algo que su creador jamás hubiese imaginado: resultaba apasionante hablar con ella.

Un cerebro profundo

La psicología y la IA tuvieron un buen punto de encuentro en la empatía de Eliza, de la misma manera en que las redes neuronales de Hopfield tuvieron un punto de encuentro con la neurociencia. ¿Cómo aprende un programa informático basado en estructuras neuronales? La respuesta vino de la principal teoría sobre el aprendizaje en el cerebro, sintetizada en la máxima que el canadiense Donald Hebb enunció en 1949: *neurons that fire together wire together* («Las neuronas que disparan juntas, se conectan»). Aquí vemos otro ejemplo de un fenómeno emergente, como la formación de patrones que estudió Turing. Cuando este mecanismo simple se aplica a una gran red da lugar a un vasto repertorio de aprendizajes en los que se cimenta la asombrosa complejidad de la inteligencia. Es el sueño de la ingeniería y de la ciencia y, en cierta medida, del arte: una regla simple capaz de explicar y sintetizar las estructuras más complejas y sofisticadas del universo.

Esta es la idea esencial de una red neuronal. Una malla lo más amplia posible, formada por distintas capas de neuronas idénticas. La combinatoria es tan grande que permite establecer circuitos capaces de codificar casi cualquier cosa. Cada patrón de activación de la red, es decir, cada conjunto de neuronas que se activan de manera simultánea, establece una representación «mental» de un objeto. Puede ser la representación de algo concreto como un animal,

o de un ente abstracto. Estas estructuras, a su vez, pueden combinarse para formar representaciones más complejas. Para poner un ejemplo matemático: la activación de un grupo de neuronas puede indicar si un número es par. La activación de otro grupo de neuronas, si un número es mayor a cien. Estos dos circuitos pueden combinarse en uno nuevo para representar los números pares, que además son mayores a cien. Una red neuronal así establece una relación unívoca entre los objetos y sus representaciones en grupos específicos de neuronas. Las neuronas que se activan cuando la red ve este objeto, por la regla de Hebb, se conectan entre ellas. Y en este entramado particular queda el recuerdo de un objeto que puede activarse y representarse de manera abstracta. En este momento, mientras leés estas páginas, se están creando nuevas conexiones entre las neuronas de tu cerebro y se están fortaleciendo otras que ya existían. Esos cambios en tu red de neuronas constituyen la manera en la que se construye el recuerdo de esta lectura. El registro de información en una red neuronal artificial opera del mismo modo.

Las redes neuronales artificiales se organizan en una estructura jerárquica de capas sucesivas, otra idea tomada del cerebro humano. En su versión más simple incluyen tres capas: una de entrada que codifica la información recibida, otra intermedia que la procesa y representa de manera más abstracta, y una de salida para dar una respuesta. Gracias al aumento del poder de cómputo del *hardware*, fue posible agregar cada vez más cantidad de capas intermedias, dando lugar a un nuevo tipo de red neuronal conocida como aprendizaje profundo, o por su nombre en inglés *deep learning*.

Articulando un número enorme de capas, este tipo de red se volvió sumamente potente y poco a poco empezó a reducir la gran brecha que la separaba de un cerebro humano. Como todas las redes, establece representaciones (también llamadas atributos). Las representaciones generadas en una capa sirven como elementos para la fase siguiente, que logra así un nivel de abstracción mayor. Esa característica las vuelve muy potentes y empieza a dotarlas de rasgos de la inteligencia humana, como el mencionado de la abstracción.

El ejemplo paradigmático, uno de los más estudiados en nuestro propio cerebro y el que sirvió como territorio de pruebas en la IA, es el de la visión. En la corteza visual hay una primera capa que detecta los bordes donde cambia la luminosidad o el color. Estos son los ladrillos básicos del sistema visual, sus primeros atributos. Luego, una segunda capa toma esta información ya procesada y empieza a combinar esos segmentos para codificar formas de geometría sencilla: un ángulo recto, un ángulo inclinado, una «T», un cuadrado... A su vez, esta capa se convierte en el insumo de la siguiente, que la recombina para procesar formas más complejas, como una cara, hasta lograr codificaciones abstractas de objetos complejos (un gato, una persona feliz, Pedro, un amanecer de invierno). El resultado de este cálculo secuencial de la red es identificar todos los atributos que hacen que un «gato» sea un gato. Eso permite que el cerebro lo reconozca sin importar si es adulto o bebé, o si está de perfil, acostado, dibujado por un pintor impresionista, arqueado, dormido, o saltando... Todas estas imágenes tan distintas corresponden al mismo concepto: tienen en común aquello que define la esencia de qué es un gato. Este trabajo de abstracción o categorización es central para la inteligencia y se resuelve de una manera relativamente sencilla. Brutal en su esfuerzo computacional y en los cientos de millones de neuronas necesarias, pero simple en su lógica y procedimiento.

El alumno supera al maestro

Las redes neuronales cambiaron la forma de aprender de las máquinas: ya no se programan con una serie de instrucciones escritas por un humano, sino que se entrenan para que vayan descubriendo los patrones de conexiones neuronales que las vuelven efectivas. En este proceso aparece un elemento que también está en la esencia del aprendizaje humano: la retroalimentación o *feedback*. Volvamos al ejemplo que ya hemos visto: una red neuronal tiene que responder si una imagen corresponde a un gato o no. Al principio sus conexiones son arbitrarias y por lo tanto su desempeño será casi azaroso.

Pero, y aquí está la clave, cada vez que reciba la indicación de que ha acertado, el patrón de conexiones que condujo hasta ese acierto se reforzará, aumentando la probabilidad de que esa respuesta se repita en situaciones similares. Por el contrario, cuando se le indique que cometió un error, las conexiones que llevaron a ese desacierto se debilitarán, generando el efecto opuesto.

Así, en un proceso laborioso, que a la velocidad de una computadora es posible en tiempos razonables, la red va aprendiendo la estructura precisa de conexiones que le permite resolver esa tarea. Pasado este entrenamiento, puede responder de manera exitosa a nuevas imágenes que nunca ha visto. En este momento vale utilizar la metáfora de que la red ha entendido lo que es una categoría. Para eso habrá formado algunas conexiones específicas que se corresponden con los atributos que debe utilizar para identificar esa categoría de manera acertada. Este ejemplo fácil de expresar, no de resolver, se extiende a casi cualquier problema que podamos asociar con la inteligencia, aun con los que son en apariencia más sofisticados. Este mecanismo es una versión simple de lo que se conoce como «aprendizaje por refuerzos» (reforzar los patrones que funcionan) y, por más elemental que parezca, está en los cimientos de la inteligencia humana y artificial.

Aparece aquí un hallazgo sorprendente: aprendiendo por su cuenta sobre la base de este proceso, el alumno (la red neuronal) puede entender el problema mejor que el maestro (el ser humano) que le ha presentado estos casos o, en otras palabras, adquirir una capacidad sobrehumana para esa tarea particular. Lo hace identificando atributos clave para resolver el problema que nosotros no contemplamos o que no podemos verbalizar. Surge así otra sorpresa: la manera en la que una red neuronal resuelve un problema puede volverse incomprensible para los humanos. La palabra «incomprensible» se usa aquí en un sentido literal. Así como un conjunto puede describirse por extensión (enumerando todos los elementos que lo componen) o por comprensión (escribiendo una regla que permite identificarlos), algo se vuelve incomprensible cuando no es posible expresar esa regla de manera verbal. Así, las máquinas pueden hacer las cosas, pero no explicarnos cómo las hacen, y pasan a ser un enigma para nosotros.

La receta para aprender a través de la retroalimentación que recién presentamos es bastante simple. El problema es que la vida está repleta de situaciones en las que nadie puede decirnos si lo que hemos hecho está bien o mal, pero de todos modos hay que aprender. Esto se resuelve, tanto en el cerebro humano como en las redes neuronales artificiales, creando una función de valor: una representación abstracta de «cuán bien se ha hecho algo». Por ejemplo, en el caso de un juego, que es su versión más sencilla, la función de valor de una posición determinada representa la probabilidad de ganar a partir de ese punto. Una jugada es buena si nos lleva a una posición mejor, es decir si aumenta esa probabilidad de ganar. Por lo tanto, en esta versión de aprendizaje por refuerzos, el algoritmo busca descubrir aquellas jugadas que mejoran la función de valor. Así, la función de valor opera como una representación interna de la retroalimentación. El algoritmo refuerza conexiones cuando aumenta la función de valor y las cambia cuando disminuye. El programa realiza este proceso de aprendizaje por sí solo, simplemente jugando. Esto no es tan raro. Muchos hemos aprendido un juego sin leer las reglas, simplemente observando, ensayando y aprendiendo a partir del éxito o del fracaso de lo que hemos hecho. El tema es que fuera de los juegos puede ser muy difícil y arbitrario establecer esta función. Aun así, la clave es que el *input* humano a la red neuronal es definir la función de valor, indicándole qué es lo que debe maximizar. Luego, el algoritmo de aprendizaje por refuerzos resuelve de manera muy efectiva esta tarea. Una vez que le decimos el qué, la IA encuentra el cómo.

Todos los juegos, el juego

Estas redes neuronales, famosas por resolver todo tipo de problemas de la vida cotidiana, no son muy distintas de las que concibió Hopfield hace más de cuatro décadas. Yann Le Cun, actual director del departamento de investigación de Meta y uno de los pioneros de la IA contemporánea, publicó dos trabajos que establecen las bases fundacionales de las redes profundas: cómo se configuran, cómo

aprenden, cómo se estructuran y cómo pueden entrenarse para usos prácticos. Esos dos trabajos son de 1989 y 1998, es decir del siglo pasado, cuando la mayoría de los que hoy interactúan con la IA ni siquiera habían nacido. ¿Por qué esta tecnología estuvo tanto tiempo en un estado semiletárgico? ¿Qué fue lo que hizo que, de repente, irrumpiera en el mundo?

Durante muchos años, las redes neuronales profundas estuvieron frenadas por la ausencia de *hardware* que estuviera a la altura del nivel de procesamiento de datos que requerían. No solo por su capacidad de cálculo sino, más importante aún, por la forma en la que procesaban información los chips más habituales en las computadoras del momento, conocidos como CPU (Central Processing Unit, o unidad central de procesamiento). Este tipo de componentes realiza cálculos a una velocidad pasmosa, pero lo hace de manera secuencial, uno después de otro. Esto resultaba una limitación importante para la industria de los videojuegos, que necesitaba controlar de manera simultánea un enorme número de píxeles en imágenes, para generar escenarios virtuales fluidos, que resultaran detallados e inmersivos. Gracias a la necesidad específica de esa industria, se popularizó a lo largo de los últimos veinte años un tipo de dispositivo computacional específicamente diseñado para el procesamiento de gráficos llamado GPU (Graphics Processing Unit, o unidad de procesamiento de gráficos), capaz de realizar un gran número de cálculos en paralelo en vez de hacerlo de manera sucesiva. En juegos como el Fortnite o el Counter-Strike, en los que la pantalla cambia a una velocidad vertiginosa, hace falta actualizar cada píxel según la dinámica del juego. Y para esto no conviene usar un gran procesador central, sino realizar en paralelo muchas instancias de una tarea más simple. Las GPU lograron aligerar la carga de trabajo del procesador central y de esta forma, mientras gran parte de lo relacionado con los gráficos se resuelve en la GPU, la CPU puede dedicarse a otro tipo de operaciones.

El funcionamiento de este tipo de *hardware*, en el que una multitud de cosas se resuelven a la vez, se parece mucho más al de nuestro cerebro, acostumbrado a realizar múltiples operaciones de forma paralela y, por eso, se presta más a los modelos utilizados

para emular la inteligencia. Las redes neuronales finalmente encontraron en las GPU el lugar donde expresarse como pez en el agua.

El otro ingrediente que hizo que la IA explotase en los últimos años es más evidente. Un espacio digital en el que se han depositado casi todos los datos humanos. La historia empezó en 1971, cuando se conectaron veintitrés computadoras a la red del Departamento de Defensa de los Estados Unidos, ARPANET, y Ray Tomlinson envió el primer correo electrónico. Unos veinte años después se configuró el protocolo de la red informática mundial (www) y unos diez años más tarde esta red acumulaba una cantidad ingente de datos de producción humana que permitió alimentar a las redes neuronales, capaces de digerirlos vorazmente para forjar su inteligencia. Ahí encontraron todos nuestros textos, las imágenes de pinturas realizadas durante miles de años, millones de cartas, mensajes, confesiones, amenazas, decisiones en todo tipo de juegos y negocios, lo que compramos y lo que vemos, nuestros secretos más privados, la expresión de dudas y certezas en la velocidad con que apretamos un botón o la lentitud con la que pasamos de una imagen a otra...

Y así fue como una necesidad específica de la industria de los videojuegos y el desarrollo de internet se convirtieron en los componentes que dieron combustible a la bomba de la IA, que había esperado paciente y silenciosa durante años a que se diesen, al fin, las condiciones necesarias para su explosión. Y explotó.

La jugada que lo cambió todo

Los nuevos mastodontes de la IA, equipados con sus numerosas capas profundas y ejecutados en redes paralelas de muchos procesadores, también se pusieron a prueba en un tablero. En este caso el juego elegido para la *batalla final* fue el go, un juego de mesa originado en China hace más de dos mil quinientos años. La compañía DeepMind presentó en 2015, un año después de haber sido

adquirida por Google, una IA entrenada sobre la base de millones de partidas humanas. Para entender la dificultad esencial del proyecto, es importante saber que el go presenta muchas más combinaciones que el ajedrez: ¡existen más posiciones posibles en un tablero de go que átomos en el universo! Y, por eso mismo, muchos especialistas creían que sería imposible que una máquina pudiera jugar competitivamente. Hasta que llegó AlphaGo.

Este programa, que entraría con Eliza y algunos otros en el panteón de las inteligencias artificiales, se encontró en su primera puesta en escena con el mayor de los desafíos; su contrincante, embajador de lo humano, era el coreano Lee Se-dol, ganador de ocho títulos mundiales. El partido se transmitió por *streaming* ante doscientos millones de personas. En la segunda partida, AlphaGo hizo un movimiento sorprendente, una jugada que ningún humano (al menos uno que jugase bien al go) hubiese hecho. Los expertos que comentaban la transmisión observaban azorados como la máquina cometía un error de principiante. Se-dol quedó perplejo, dejó la sala y necesitó quince minutos para intentar entender lo que pasaba. Pero AlphaGo demostró que ese supuesto error combinado con otras ideas que nadie había considerado era en realidad un gran acierto, y ganó la partida. La jugada resultó absolutamente revolucionaria, original y creativa, y cambió la manera en la que los humanos abordaron el juego a partir de ese momento. Nos regaló una idea nueva que no se le había ocurrido a ningún jugador en los casi tres mil años de historia del go. Las máquinas por primera vez parecían listas para superarnos, incluso en la esfera más humana: la creatividad.

De la misma manera en la que un pintor o un escritor pueden introducir una nueva manera de retratar o de narrar, AlphaGo introduce una innovación que cambia y enriquece la manera en que jugamos nosotros, los seres humanos. Fue, quizá, el primer legado introducido por una máquina en la cadena de una cultura milenaria. Pasa la gente y los programas, pero aquella movida que introdujo AlphaGo ha quedado en el repertorio. Hoy, los maestros en China y Japón enseñan esta jugada en las clases para principiantes de go.

AlphaGo era extraordinaria por su capacidad de concebir nuevas ideas estudiando todo el repertorio de partidas humanas. Pero

en 2017, AlphaZero, el sucesor de AlphaGo, dio un paso más y aprendió a jugar tanto al ajedrez como al go sin que nadie le aportara un solo concepto estratégico ni le enseñara una partida. Aprendió jugando contra sí misma.

¿Cómo lograr algo tan asombroso? La máquina comienza jugando contra una copia de sí misma, y la clave está en que solo se le permite a una de las dos copias revisar su modelo de juego. Después de miles de partidas, la que puede aprender comienza a vencer a la otra de manera sostenida. En ese momento se realiza un nuevo clon de esta versión mejorada y se repite el proceso. Así, en esta repetición alucinante de juegos simulados, en medio de la noche cibernética, un programa jugando contra sí mismo genera un conocimiento exponencial. Con eso, la inteligencia artificial estaba lista para el próximo gran desafío.

2

Una nueva era

Leer la mente

Andre Agassi y Boris Becker, dos de los grandes jugadores de la historia del tenis, entablaron una rivalidad legendaria. Sus estilos de juego eran opuestos: mientras Agassi era conocido por su habilidad en el fondo de la cancha, su agresividad y su capacidad para devolver golpes desde cualquier posición, Becker se destacaba por su saque potente y su habilidad en la red. Sin embargo, el famoso saque de Becker nunca fue muy efectivo contra Agassi. En 2009, cuando el tenista de Las Vegas publicó su autobiografía, se entendió por qué. Allí revela un secreto que mantuvo oculto durante años. Agassi descubrió que Becker, sin darse cuenta, hacía un movimiento con la lengua que delataba el tipo de saque que estaba a punto de efectuar. Gracias a ese gesto o atributo que había pasado inadvertido para el resto de sus rivales y para millones de televidentes, Agassi logró descifrar ese aspecto clave del juego de su rival. Así lo cuenta en su libro: «Becker era un poco obvio. Hacía un movimiento recurrente con la lengua cuando se balanceaba para ejecutar el saque: si cerraba la boca, el saque iba al centro de la pista; si deslizaba la lengua hacia un costado, entonces seguramente realizaba un saque abierto». Agassi tuvo que decidir con cuidado cómo usar su hallazgo. «La parte más difícil fue que no se diera cuenta en la cancha de que yo sabía lo que hacía con la lengua. Así que tuve que resistir la tentación de leer sus saques continuamente y elegir el momento en el que usar esa información», confesó. Agassi tenía,

en el mundo del tenis, una superinteligencia que le permitía detectar rasgos casi imperceptibles para predecir la dirección de un saque. Una red neuronal funciona de la misma manera: detecta atributos que le permiten identificar si una imagen es o no la de un gato, si hay un tumor en la imagen de un pulmón o qué emoción en particular expresa la voz de una persona. Estos atributos permiten sacar conclusiones y tomar buenas decisiones en dominios muy específicos. Como a Agassi, nadie le enseña a una red neuronal cuál es el mejor atributo para poder predecir algo. Tiene que descubrirlo a partir de una pila abismal de datos.

Leer los atributos del adversario permite anticipar las acciones y movimientos del oponente y ajustar las estrategias y tácticas en consecuencia. Tomemos uno de los ejemplos más simples, el célebre juego piedra, papel o tijera. Todos lo percibimos como un juego de azar pero, a la vez, entendemos que se trata de intentar leer la mente del rival, de adivinar qué es lo que va a elegir a partir de sus gestos o de su historial de elecciones previas. Se trata de encontrar patrones y atributos. Si un programa fuese incapaz de leernos la mente, o si nuestras elecciones fuesen completamente azarosas, no podría vencer de manera sistemática a una persona. Pero resulta que nuestras decisiones al azar no son tan azarosas después de todo. Somos más previsibles de lo que suponemos, y elegimos de acuerdo con un mecanismo inconsciente lleno de regularidades detectables —como los movimientos de la lengua de Becker— aun cuando otras personas e incluso nosotros mismos seamos incapaces de descubrirlas. Es decir, dejamos todo tipo de trazas de nuestras elecciones, que una red neuronal puede utilizar como atributos para inferir nuestra próxima jugada.

Por lo tanto, no debería sorprendernos que en 2020, en la universidad china de Zhejiang, se diseñara una IA que detecta estos patrones ocultos con tal precisión que acabó venciendo al 95 por ciento de las personas con las que se enfrentó en partidas de piedra, papel o tijera a trescientas rondas. Como en el go, el mejor jugador del mundo de piedra, papel o tijera también es una IA. Precisamente porque aprende a detectar patrones que para la mayoría de nosotros son imperceptibles.

Las IA resuelven todo tipo de problemas mejor que cualquier humano porque acceden a atributos que son efectivos para el problema que las máquinas intentan resolver. A nosotros esto nos cuesta más por la dificultad que supone identificar los rasgos más relevantes entre tantos posibles. Este es el caso del saque de Becker. Los datos están ahí, accesibles para todos, pero nuestra atención limitada hace que casi nadie, salvo un genio del juego, repare en ellos.

Otras veces, hay una restricción estructural más evidente. Nuestros dispositivos sensoriales tienen limitaciones importantes, de las que no somos conscientes. Tomemos como ejemplo el oído: el rango de audición humana detecta los sonidos con frecuencias entre 20 y 20 000 Hz. Por lo tanto, una máquina, al igual que algunos animales, puede oír cosas que nosotros no percibimos. Tenemos también una limitación más profunda: la realidad puede presentar facetas cuya existencia desconocemos por completo. Imaginemos esto: si todos fuéramos sordos, ¿cómo sabríamos de la existencia del sonido? ¿Cuántas otras propiedades del mundo se nos estarán escapando, simplemente porque no tenemos los sensores para detectarlas? En definitiva, si tomamos como ejemplo el procesamiento de imágenes, las redes neuronales logran ver cosas que el ojo humano no percibe, porque tienen ventajas significativas en términos de memoria, representación multidimensional, procesamiento simultáneo y adaptación.

Pensar es olvidar diferencias

Así sintetiza Jorge Luis Borges un rasgo fundamental de la inteligencia en la historia de un tal Funes, quien, tras sufrir un accidente, ha adquirido una memoria prodigiosa que lo obliga a recordar todos los detalles de su vida y del mundo que lo rodea. A priori, esto puede parecer propio de un genio, pero Borges esboza una tesis sobre la inteligencia y muestra que, por el contrario, esta memoria tan detallada entorpece uno de sus rasgos fundamentales. De cada percepción, Funes hace una característica única: «No solo le costaba comprender que el símbolo genérico "perro" abarcara tantos indi-

viduos dispares de diversos tamaños y diversa forma; le molestaba que el perro de las tres y catorce (visto de perfil) tuviera el mismo nombre que el perro de las tres y cuarto (visto de frente)».

La historia de Funes exhibe uno de los principales peligros que implica la abundancia de datos. Entender algo requiere identificar los atributos para comprender qué hace a un objeto, pero también saber ignorar aquellos que son irrelevantes. Para identificar un perro hace falta ver que ladra, que tiene orejas y cuatro patas, sin distraerse con el color, la altura, el pelaje u otros detalles de un perro en particular. La capacidad de abstraer y generalizar es, como afirma Borges, un aspecto central del pensamiento.

Y en esto también las máquinas tienen todo lo que necesitan para aventajarnos: ven cosas que nosotros no podemos ver. También nos superan ampliamente en la capacidad de identificar, entre millones de características, cuáles son las más eficientes para resolver un problema, y cuáles son las irrelevantes, las que conviene ignorar. Eso es exactamente en lo que las IA son excelsas.

Por eso, una IA puede, en poquísimo tiempo (pero con muchos datos), determinar con mayor precisión si una radiografía muestra algún signo de patología que un médico especializado que ha estudiado durante años. Esto es gracias a esa ventaja fundamental en la capacidad de analizar todas las combinaciones posibles y ver cuáles resultan más informativas. Sobre esta base logran, en algunos dominios, una *performance* sobrehumana en términos de precisión, velocidad y alcance.

De la capacidad de entender a la habilidad de crear

Una vez que una red neuronal logra identificar todas las características que le permiten reconocer algo o a alguien en una imagen, por ejemplo, un gato, la torre Eiffel, o a Barack Obama, se abre una posibilidad inesperada. ¿Qué pasaría si se invirtiera el proceso, poniendo la red neuronal patas arriba? En vez de darle una imagen para que, basándose en sus atributos, determine si eso es un gato o no, le pedimos que, a partir de todo lo que sabe, produzca una

imagen que ella misma categorizaría como un gato. La manera en que esto se logra es invirtiendo el flujo de información, utilizando la capa de salida como entrada y viceversa. Se activa la neurona de un gato en la capa más abstracta y esta capa va proyectando hacia atrás, activando primero los atributos de mayor nivel y progresivamente los de menor nivel hasta llegar a los trazos, sombras y ángulos de una imagen que cumple con todos los atributos de un gato genérico. Esta imagen que se produce en las primeras capas de la red es como un sueño. Es una invención nueva: se ha creado un gato que no existía hasta ese momento.

Si bien las redes que vimos en las secciones anteriores ya mostraban cierta capacidad creativa, el mecanismo de inversión que acabamos de presentar cambia abruptamente sus posibilidades. No solo se puede inventar una jugada, sino una imagen, un sonido o una frase. Aquellas cosas que son el combustible de nuestra percepción o el resultado de nuestra imaginación.

El gato que produce la red neuronal no es ninguno de los gatos que vio en su entrenamiento. Es un animal nuevo y único, que cumple con todos los requisitos necesarios para pertenecer a esa categoría. En otras palabras, cada uso es esencialmente un proceso creativo. A este nuevo tipo de redes neuronales se las llamó redes generativas.

La idea era buena pero los resultados iniciales fueron bastante pobres. Los primeros intentos generativos estaban llenos de imperfecciones y engañar al ojo humano no es una tarea sencilla. La solución a este problema llegó de la mano de un informático e ingeniero estadounidense llamado Ian J. Goodfellow, que encontró una manera genial de refinar este proceso. La idea se le ocurrió una noche de 2014, en un bar de la ciudad de Montreal, mientras charlaba con su supervisor de doctorado, Yoshua Bengio, sobre cómo lograr que las redes neuronales produjeran creaciones verosímiles.

¿Por qué no poner dos redes neuronales a competir entre sí para que aprendan de sus errores? Enfrentar a una red generativa —que produce objetos similares a los usados en su entrenamiento— con una discriminadora, que usa su habilidad especialmente

afinada para detectar si el material proporcionado es falso o no. Así nació lo que se conoce como Redes Generativas Adversariales o *Generative Adversarial Networks* (GAN). A la discriminadora se le entrega una mezcla de datos reales y otros generados, y debe evaluar e identificar los falsos. Ambas redes reciben *feedback* en cada iteración: si la generadora es descubierta, debilita las conexiones neuronales que llevaron a ese intento fallido. Por el contrario, si logra superar el filtro de la discriminadora, las conexiones que lo han conseguido se fortalecen. Al otro lado, la discriminadora también va aprendiendo de sus aciertos y errores para hacer cada vez mejor su trabajo.

Cuando comienza la contienda entre ambas redes, la generadora funciona de manera bastante burda y los objetos inventados no se parecen mucho a los reales. Por eso, si su objetivo es generar un gato, al principio tendrá alteraciones o problemas evidentes, y esas fallas serán fácilmente detectadas por la discriminadora. Con el paso del tiempo, la red generadora va logrando objetos cada vez más realistas y la red discriminadora empieza a perder algunas batallas y necesita refinar su capacidad para seguir compitiendo. Finalmente, llega un punto en que la generadora le gana la partida a la discriminadora. Empieza a producir objetos tan similares a los reales que ni siquiera una poderosa red neuronal entrenada a tal efecto puede detectar que son falsos. Llegado este punto también es posible engañar al ojo humano.

Este enfoque tiene además otra virtud: a diferencia de lo que ocurre con las redes discriminadoras y con las generadoras sin un oponente, las GAN pueden entrenarse con una cantidad mucho más pequeña de datos. La capacidad de cada una de hacer bien su parte depende ahora menos del material usado para el entrenamiento y más de esta competencia entre ambas. Una vez más, encontramos que las inteligencias artificiales pueden aprender y alcanzar niveles superlativos, sin requerir de la intervención o la habilidad humana.

El mundo visual ha sido, como el juego, el terreno de prueba y espacio de exhibición de los mayores avances en inteligencias artificiales, tanto en el reconocimiento como en la generación. Por eso hemos recurrido a muchos ejemplos de capas profundas en el

mundo de la percepción visual. Evidentemente la misma estrategia puede aplicarse para la generación de otro tipo de datos, ya sea sonido para clonar una voz o imagen en movimiento para simular video. Aunque todo parezca igual, no lo es. Porque, como veremos pronto, algo muy especial ocurre cuando aplicamos esta misma idea a la generación de lenguaje. Para verlo, antes necesitamos resolver otro problema.

Los secretos más íntimos del lenguaje

Como vimos, la capacidad de hacer cálculos en paralelo de las GPU dio una nueva vida a las IA, sobre todo en el mundo de la imagen, en el que la información se presenta al unísono. Pero reconocer lenguaje es una tarea bien distinta. Porque el lenguaje, como la música, sucede en el tiempo. Decía la filósofa Susanne Langer que la música es el laboratorio para sentir en el tiempo y algo parecido sucede con las ideas y el lenguaje; la información se expresa de manera secuencial, una palabra detrás de otra, y el sentido de muchas palabras solo puede entenderse de acuerdo con el contexto en que se utilizan: qué viene antes y qué después. Por eso la comprensión del lenguaje está plagada de contexto, de cosas no dichas que se presuponen, y de expectativas que se construyen en el tiempo presente con la información del pasado. Podemos entenderlo con el ejemplo de nuestro amigo Juli Garbulsky: «Es increíble cómo la gente se sorprende cuando una frase no termina exactamente como ellos lechuga». «Lechuga» funciona en esa frase como una nota desafinada, una entrada que está fuera de las expectativas que el cerebro ha ido construyendo, a la velocidad vertiginosa en la que se suceden las palabras en una oración. El lenguaje sucede en el tiempo y no en el espacio. Si en un párrafo mencionamos a una persona específica, habitualmente evitamos ser redundantes y sobreentendemos que el lector recuerda con facilidad lo que ha leído algunos segundos o líneas antes.

Del mismo modo, si nos proponemos que una máquina realice esta tarea lingüística necesitaremos recurrir a referencias que

están en el pasado para comprender lo que significan las palabras. Para resolver este problema, a principios de este siglo se utilizó un tipo diferente de red neuronal llamado «redes neuronales recurrentes» (RNN), que incorporan un mecanismo de memoria. Sin embargo, los avances fueron lentos y tortuosos. Y pronto quedó claro que, en este caso, emular al cerebro no era el camino.

La solución llegó en 2017, de la mano de un artículo publicado por investigadores de la Universidad de Toronto, financiado por Google, simpáticamente titulado «Attention is all you need», en alusión a la célebre canción de Los Beatles. Y como ya había hecho la mismísima banda de Liverpool en el mundo de la música, este artículo dio comienzo a una nueva era. Lo curioso es que la idea que iba a cambiar el mundo era relativamente simple: para entender una frase no es necesario revisar todo el contexto, sino elegir bien a qué datos o conceptos mencionados antes es importante prestar atención. «Atención», ahí está la clave. Es todo lo que se necesita. En este artículo seminal se introdujo una nueva arquitectura llamada «transformer», que incorpora justamente un algoritmo para decidir cuánto peso darle a diferentes palabras o elementos de la secuencia o, dicho de otro modo, a qué prestar más atención.

Los transformers funcionan en un ensamble de dos redes profundas, una codificadora y otra decodificadora. La codificadora recibe como entrada una frase y la analiza en múltiples pasos, decidiendo cuál es la información más relevante que se guardará en la memoria para la capa siguiente. Así, logra entender mejor la oración y cómo las palabras se relacionan unas con otras. El salto de calidad en la comprensión del lenguaje natural resultó asombroso, hasta el punto de que soluciones que solo hace unos años podían parecer futuristas, como Siri o Alexa, hoy parecen increíblemente precarias. Una vez procesada la entrada de esta manera, la codificadora le pasa la información a la decodificadora, que es una red generativa que fue entrenada para producir frases verosímiles que resulten indiscernibles de las que formularía un ser humano como respuesta a la frase previa. El objetivo de este artículo era mejorar la traducción automática entre idiomas y ni sus autores ni Google vislumbraron el impacto descomunal que tendría en el futuro. Con

este hallazgo, se completaba la última pieza que faltaba para el boom actual de la IA.

Los que sí detectaron rápidamente el potencial de este nuevo tipo de arquitectura para una red neuronal fueron los científicos de OpenAI, una organización sin fines de lucro fundada casi un año antes por algunos de los emprendedores más brillantes de Silicon Valley, incluyendo a Elon Musk, Peter Thiel, Reid Hoffman y Sam Altman. Como su nombre indica, la idea detrás de esta iniciativa y este equipo de fundadores tan prominentes era dar transparencia y apertura a las investigaciones más avanzadas en IA. La mayoría de sus creadores tenían la convicción de que los riesgos de la IA eran considerables, y que una tecnología tan poderosa no podía estar en manos de una sola empresa privada. Tampoco podía verse sujeta a los potenciales incentivos perversos que genera la maximización del beneficio económico que mueve a las compañías con fines de lucro. ¿Por qué hablamos en pasado? Porque todo eso cambió un año después, en 2019, cuando OpenAI decidió convertirse en una empresa. La razón esgrimida fue que generar y distribuir ganancias entre sus empleados y accionistas era la única manera de competir con los gigantes tecnológicos como Google, Facebook y Amazon por los dos recursos más necesarios para un proyecto de estas características: el escasísimo talento especializado y el abundante dinero de los fondos de inversión. En aquel momento, descontento con el rumbo que tomaban las cosas, Elon Musk intentó tomar el control de OpenAI, y al no lograrlo se fue dando un portazo.

El cerebro científico del proyecto, Ilya Sutskever, seguramente pase a la historia, para bien o para mal, como uno de los artífices detrás de la generación de inteligencias artificiales sumamente poderosas. Envalentonado por el potencial de los transformers, se propuso hacer un experimento: ¿Qué pasaría si hiciéramos una red neuronal Generativa, Preentrenada y basada en Tansformers? Basta unir las iniciales para ver que el experimento fue exitoso, así nace GPT.

En marzo de 2023, durante una entrevista con Forbes, Sutskever contó en qué medida el descubrimiento de los transformers lo cambió todo, y recordó la rapidez con la que aplicaron aquella

idea revolucionaria: «Nuestras redes neuronales no estaban preparadas para la tarea [de predecir la siguiente palabra]. Estábamos usando redes neuronales recurrentes. Cuando salió el transformer, [...] ya estaba claro para mí, para nosotros, que los transformers resolvían las limitaciones de las redes neuronales recurrentes para aprender relaciones lejanas. Y así, el muy incipiente esfuerzo de GPT [...] empezó a funcionar mejor, se hizo más grande y más grande. Y eso esencialmente condujo a donde estamos hoy». Ahí está, en pocas frases y en primera persona, la historia reciente de la IA.

La meta de GPT era entrenar un transformer decodificador utilizando un corpus de texto descomunalmente grande, de producciones humanas agregadas durante miles de años. El método, como casi todo lo que hemos ido viendo, no fue muy sofisticado. Tomar una frase, quitar una palabra, y mejorar repetitivamente la capacidad de predecir qué término era el que faltaba. Así se volvió increíblemente efectiva para entender qué palabra va con cuál y, al captar de manera tan profunda la relación entre ellas, adquirió un conocimiento equivalente a entender la gramática del lenguaje: tanto la morfología (qué clase de palabras hay) como la sintaxis (cómo se estructuran y se ordenan). Justamente, fue el algoritmo de atención de los transformers el que le permitió disponer del contexto necesario de cada vocablo en la memoria para lograr este objetivo. Y esto se hizo no para uno, sino para al menos treinta idiomas diferentes.

Entendiendo de esta manera la lógica profunda que subyace detrás de la lengua, GPT puede construir frases increíblemente humanas, prescindiendo de la semántica (saber qué significa cada palabra). Dicho de otra manera, ha aprendido a hablar con un estilo increíblemente humano y a decir cosas interesantes y de gran trascendencia, sin tener la menor idea de lo que está diciendo.

Una vez entrenada de este modo, el siguiente paso era sencillo: usar ese conocimiento para construir respuestas, prediciendo cuál es la palabra más probable que un ser humano usaría a continuación en cada secuencia. Si, por ejemplo, partimos de la frase «Quiero un sándwich de» y analizamos todo lo que han dicho los hispanohablantes en la historia después de eso, podría predecir «jamón» o «queso», y elegir cuál de estas opciones prefiere de acuerdo con el contexto general. A su vez, la decisión que tome crea un nuevo contexto que afectará las elecciones siguientes.

El primer modelo de GPT se lanzó a mediados de 2018. Contaba con 120 millones de parámetros, valores numéricos que podían ajustarse como resultado del entrenamiento. Y se había entrenado con 4 GB de texto, menos información de la que almacenaba una tarjeta de memoria de una cámara de fotos. A partir de ahí, los modelos no pararon de expandirse. Su sucesor, GPT-2, vio la luz en febrero de 2019 y contaba con 1500 millones de parámetros, entrenados sobre la base de ocho millones de páginas web. Luego vino GPT-3, elevando más de cien veces la capacidad de su antecesor: 175.000 millones de parámetros para procesar 45 TB de datos. Y la versión más reciente, mientras escribimos este libro, es GPT-4, presentada en marzo de 2023. Haciendo honor a que OpenAI hace IA pero ya no es tan «open», el número de parámetros no fue revelado, pero se estima en... ¡1.76 billones! Esto es un uno seguido de doce ceros. Y el volumen de datos usado para entrenarlo se estima en el orden de un petabyte. Para quien nunca haya escuchado esa expresión, esto representa un poco más de un millón de GB.

Este descomunal crecimiento, similar al que fueron teniendo modelos competidores generados por Google, Facebook y otros, dio lugar a un nuevo tipo de IA. A las redes neuronales basadas en transformers, entrenadas con enormes volúmenes de texto para producir lenguaje se los bautizó como LLM (Large Language Models), es decir, Grandes Modelos de Lenguaje. Y parece que el tamaño en esto sí importa. Porque estos nuevos modelos comenzaron a mostrar resultados completamente sorprendentes, ¡incluso para sus propios creadores!

El lenguaje es el sustrato del pensamiento

Con el aumento de escala de los LLM, se abrieron puertas fascinantes e inesperadas. Podemos pensar qué sucede con la adquisición

del lenguaje en el desarrollo de un bebé. Aun cuando en los primeros meses logra aprendizajes extraordinarios, todo ese proceso cognitivo adquiere una progresión explosiva cuando consigue combinar arbitrariamente todas sus facultades gracias al uso del lenguaje. Por eso, a ningún padre o a ninguna madre se le escapa que, cuando su hijo empieza a hablar, hay todo un universo nuevo que se abre y el vínculo cambia de manera profunda, impulsado por la amplia ventana de posibilidades que dan las palabras. El uso del lenguaje permite saber a los padres por qué llora su hijo y qué le duele, y sienten una enorme y grata sorpresa cuando éste argumenta por primera vez por qué quiere hacer algo, o cuando expresa sus dudas, anhelos, miedos o sueños. Los humanos no somos mejores que el resto de los animales en el reconocimiento de objetos. Nuestra gran singularidad está en el vínculo con el lenguaje. Por eso, los LLM generan una revolución en la IA similar a la que ocurre en la inteligencia humana cuando un niño comienza a balbucear las primeras palabras.

Es que el lenguaje es la materia de la que está hecho el pensamiento humano. Cuando una IA aprende a generar lenguaje, en cierta manera está aprendiendo a pensar, aun cuando por ahora no haya en la máquina un ente que sea sujeto de ese pensamiento. Sin saber nada del significado de las palabras, puede armar un discurso interesante, profundo y coherente, aun cuando en realidad el programa no tiene ni idea de lo que está diciendo. Somos contemporáneos de esa transformación y por eso, en algún punto, también somos ese padre o esa madre que es testigo fascinado de cómo un niño comienza a pensar al articular las primeras palabras. Así como nuestros hijos dicen cosas sorprendentes, que a nosotros nunca se nos hubiesen ocurrido, las IA a veces también encuentran atributos que les permiten superarnos en algunos aspectos de la lengua.

Hace tiempo que venimos registrando avances en el procesamiento del lenguaje natural y del entendimiento de la voz. De hecho, manejar un dispositivo con comandos verbales es posible hace ya más de una década. Pero la experiencia de uso resultaba bastante limitada: muchas veces confunden lo que les decimos y el repertorio de instrucciones que comprenden es bastante acotado. La

primera sorpresa con los LLM, especialmente a partir de GPT-3.5, es que muestran un grado de sutileza y profundidad en el entendimiento muy por encima de sus antecesores. Parecen entender la ironía, el humor y otras propiedades sutiles que expresamos en un texto.

Las posibilidades que emergen en la historia de la IA y que hemos ido enumerando alternan entre cambios cualitativos y cuantitativos. Cantidad y calidad muchas veces se entremezclan, como explicó Friedrich Hegel en su libro *Ciencia de la Lógica*: «Cuando hablamos de un crecimiento o una destrucción, siempre imaginamos un crecimiento o desaparición gradual. Sin embargo, hemos visto casos en los que la alteración de la existencia implica no solo una transición de una proporción a otra, sino también una transición, mediante un salto repentino, a algo cualitativamente diferente; una interrupción de un proceso gradual, que difiere cualitativamente del estado anterior». Este fue el mecanismo mediante el cual lo cuantitativo se transformó en cualitativo que propuso Hegel y que luego Karl Marx usó para explicar las grandes transformaciones sociales. Las IA no están exentas de esta regla. Los LLM pueden realizar tareas complejas en poquísimo tiempo, comparado con los meses o años que le llevarían a las pocas personas expertas capaces de realizarlas. La velocidad de estos procesos dará lugar a cambios cualitativos en el valor de una idea, en el número de creaciones que podremos sintetizar y en nuestro vínculo con la educación y con el trabajo.

¿Qué pasaría si un día, en ocasión de una de esas preguntas retóricas que le hacemos a un perro, éste decidiera responder de golpe? Si nos dijera, al preguntarle si quiere salir, que lo agradece, pero que prefiere no hacerlo, que está algo melancólico y prefiere quedarse en casa hasta que la tristeza amaine. Quedaríamos atónitos y al segundo fascinados por la enorme cantidad de cosas que a partir de ese momento podríamos compartir con él, y es que el lenguaje es la ventana de acceso más privilegiada a la mente de otra persona o de cualquier otro ser.

Este ejercicio mental sencillo y hogareño nos permite intuir la dimensión del salto cualitativo que implica que las redes neurona-

les hablen nuestro idioma. Se crea un puente empático que nos da una cercanía nueva y profunda. Al mismo tiempo, nos ofrece un canal de comunicación mucho más amplio, en el que el repertorio de cosas que les podemos pedir se expande en proporción al casi infinito poder del lenguaje y su capacidad combinatoria. Por eso, las redes neuronales que poblaron el mundo sin mayor trascendencia durante décadas, de repente se volvieron celebridades en el momento mismo en el que, transformers mediante, adquirieron el lenguaje. Y, de un día para otro, se pusieron a hablar.

En esa riqueza radica el corazón de lo humano: el lenguaje nos permite contar historias, tener objetivos, creencias, valores, sueños y deseos, pero también tener ideas y poder cambiarlas. Gracias a la manera en que se articula y organiza el pensamiento, nos posibilita entender y describir cada rincón del universo. Bienvenidas las máquinas a este lugar privilegiado que la inteligencia hasta hoy solo reservaba al ser humano.

3

El punto de llegada es un nuevo punto de partida

Saber qué hacer cuando no sabemos qué hacer

En los capítulos anteriores, hemos recorrido el camino que nos llevó a desarrollar máquinas capaces de abstraer, de calcular, de generar ideas propias y originales, de concebir objetos y, en última instancia, de conversar. Todas esas son características propias de la inteligencia. Pero no son las únicas.

Otro rasgo de la inteligencia es, justamente, cuestionarse sobre la naturaleza misma de la inteligencia. Por eso, a lo largo de la historia humana, ha habido tantos intentos de definir de manera exhaustiva y precisa de qué se trata. Pero cada uno de estos intentos ha terminado en un callejón sin salida: «No somos lo suficientemente inteligentes como para definir inteligencia», dice con mucho criterio nuestro amigo físico Gerry Garbulsky.

La inteligencia no se limita únicamente a la capacidad de razonamiento lógico o al procesamiento de información, sino que también involucra, entre otros, aspectos emocionales y sociales, o la relación misma con el cuerpo. ¿O acaso queda alguna duda de que una gran coreografía o una jugada extraordinaria de fútbol son grandes despliegues de inteligencia humana? La gran diversidad de sus manifestaciones hace que sea complejo elaborar una definición acabada y exhaustiva.

Una célebre definición de inteligencia popularmente atribuida al psicólogo y epistemólogo suizo Jean Piaget es: el arte de saber qué hacer cuando no sabemos qué hacer. Es decir, la capacidad de

encontrar una solución a una situación compleja que no hemos vivido antes. Según esta definición, la inteligencia es una herramienta versátil y flexible, que nos permite adaptarnos a nuestra realidad, a nuestro entorno, y resolver los problemas que aparecen. Nosotros, para los propósitos de este libro, adoptaremos esa visión amplia. Sabemos que siempre encontraremos excepciones que nos dejarán insatisfechos y con sabor a poco. Aceptamos que el concepto es más rico de lo que podemos capturar en una definición y, ante eso, nos conformamos con acercarnos al término para, en ese movimiento, explorar sus posibilidades.

También es difícil distinguir la inteligencia de otros conceptos, como por ejemplo el de la cultura. Veámoslo a través del siguiente experimento. Un grupo de monos está en una jaula en cuyo suelo hay algunas cajas y desde lo alto de la cual cuelga un racimo de bananas. Los monos descubren, haciendo gala de un tipo de inteligencia que compartimos con otros primates, que pueden apilar estas cajas para llegar a la parte superior de la jaula y alcanzar las bananas. Pero en este experimento hay un truco algo malvado: cada vez que uno de ellos está cerca de agarrar la fruta, una repentina ducha de agua helada los moja a todos. Tras varias repeticiones, cuando alguno de los monos comienza a apilar las cajas para trepar en busca de bananas, los otros lo frenan con cierta violencia porque han aprendido que, si llega alto, terminarán empapados. Cuando este rasgo «cultural» ya está asentado, se reemplaza a uno de los monos de la jaula por uno nuevo que, naturalmente, comienza a apilar las cajas. Apenas comienza, es bruscamente frenado por los demás. Tras algunos intentos, sin entender la razón del castigo, acepta la regla y aprende a no intentar subir. Cuando eso sucede, se suma otro mono nuevo y se repite el proceso. Así se continúa paso a paso hasta que llega un punto en el que ninguno de los monos en la jaula ha experimentado nunca el agua helada. Sin embargo, todos respetan y hacen respetar la regla de que las cajas no deben apilarse. Esa norma cultural es compartida por el grupo, a pesar de que ninguno sepa el porqué.

En este ejemplo vemos muchas manifestaciones de la inteligencia: la idea inicial de apilar las cajas para llegar hasta arriba; la

capacidad de concluir que, aun siendo posible subir, conviene no hacerlo; la posibilidad de aprender esa regla y de enseñarla. Vemos también que, pasado un tiempo, la decisión de reprimir el deseo de alcanzar el racimo de bananas, que puede parecer un reflejo de inteligencia, en realidad quizá no lo sea. Es, más bien, el reflejo de la inteligencia de generaciones anteriores expresada en forma de cultura. Si en algún momento el mecanismo del agua fría fuera desactivado, los monos no cambarían su comportamiento. Este es un ejemplo de laboratorio del concepto de «servidumbre voluntaria» que introdujo el filósofo francés Étienne de la Boétie hace más de quinientos años. Este concepto nos servirá como guía en los capítulos que siguen para evitar que nosotros mismos, primates humanos, inmersos en un mundo de algoritmos, y a los que cada tanto riegan con agua helada, dejemos de apilar cajas sin saber por qué.

Solemos entender la capacidad de aprender como uno de los rasgos distintivos de la inteligencia humana. Pero en realidad está extendida en el mundo animal y hasta organismos muy sencillos la tienen. De hecho, los mecanismos del aprendizaje se han descubierto en gran medida estudiando al *Caenorhabditis elegans*, un gusano de menos de un milímetro de longitud, que tiene un 80 por ciento de genes homólogos a los del ser humano. El análisis exhaustivo de la babosa de mar *Aplysia californica* llevó también al premio Nobel Eric Kandel a descubrir la mecánica molecular y celular de la memoria. Aprender, entonces, no es un rasgo propio de la especie humana, aunque muchas veces lo creamos. En cambio, lo que es menos común es la habilidad que tiene nuestra especie para enseñar, para propagar el conocimiento como un virus contagioso. Puede decirse que más que una habilidad, es una vocación irrefrenable: la voracidad por compartir lo que hemos hecho o lo que conocemos es una pulsión tan innata como beber o buscar alimento. No nacemos sabiendo enseñar, como tampoco nacemos sabiendo caminar o hablar, pero hay un programa de desarrollo que promueve y estimula esta habilidad y que se pone en marcha en las primeras semanas de vida. Y el resultado de esta vocación por compartir lo descubierto es el vasto repertorio de la cultura humana,

con obras que van desde las esculturas de Miguel Ángel, a las sinfonías de Mozart, los goles de Maradona o el teorema de Pitágoras. Cada una de estas creaciones es consecuencia del proceso reiterado de aplicar la creatividad e inteligencia de una época sobre la cultura y conceptos generados anteriormente, para alcanzar estratos cada vez más sofisticados.

Podemos distinguir la inteligencia del aprendizaje y de la cultura con un ejercicio mental. ¿Qué le ocurriría a una persona adulta nacida hace diez mil años si la trajéramos al presente? A pesar de tener un cerebro morfológicamente igual al nuestro, no podría acomodarse a la vida moderna: no hablaría la lengua, no sabría cruzar la calle ni subir a un ascensor, no entendería qué significa que alguien le guiñe un ojo. Esa persona parece desprovista de inteligencia, pero no es así. Tiene una laguna cultural. Esto queda más claro si pensamos que un bebé nacido hace diez mil años, transportado al presente y criado aquí sería indistinguible de cualquiera de nosotros y viviría una vida normal.

Nuestro cerebro es prácticamente igual al de nuestros antecesores de hace diez mil años, lo que nos diferencia es una inteligencia cultivada y agregada en una historia educativa y cultural. La inteligencia, entonces, es acumulativa. Esto es cierto para todas las inteligencias, artificiales o humanas. Muchas IA hoy nos sorprenden no porque sean más inteligentes, sino simplemente porque son más cultas. Aplican la misma inteligencia a un repertorio más sofisticado de ideas, cocinado en ciclos de inteligencia sucesivos que, en silicio, en vez de producirse en miles de años, se realizan en pocos días.

¿Cómo se construye una inteligencia?

Cuando Alan Turing dio los primeros pasos en la creación de la IA, tomó un camino bastante pragmático. Se propuso emular la inteligencia que mejor conocía: la suya. Observó y buscó imitar sus propios razonamientos, y fue pionero en generar un laboratorio de autoobservación para explorar las posibilidades y los límites de la mente humana. Por eso la IA fue, en su origen, un lugar de con-

fluencia entre la psicología y la filosofía de la mente. Replicar la inteligencia era una forma de entenderla y entenderla era una forma de replicarla. Este parece el camino más simple e intuitivo para emular cualquier habilidad. Por ejemplo, en las primeras concepciones de máquinas voladoras, se usaban alas ligeras que subían y bajaban a gran velocidad para intentar lograr sostenerse. Buscábamos volar imitando el mejor ejemplo de una «máquina» voladora que conocíamos: los pájaros. De igual forma, en los orígenes de la IA buscamos imitar los modos y capacidades de la especie más inteligente que conocíamos: la humana.

Sin embargo, la historia de la aeronáutica pronto toma un camino muy distinto. Nuestra comprensión de la física de la sustentación, la propulsión, la gravedad y la aerodinámica nos permitió desarrollar artefactos capaces de volar de maneras que poco tienen que ver con el vuelo de las aves o los insectos. Independizarnos del modo en que vuelan otros organismos permitió que nuestras máquinas voladoras alcanzaran mayor velocidad, transportaran significativamente más carga y gozaran de una autonomía más amplia.

De la misma manera, hemos resuelto el problema de la movilidad sin imitar la caminata ni el galope de un caballo. La rueda nos ha resultado un medio mucho más conveniente para optimizar nuestro transporte y hemos fabricado carros, trenes, motos y automóviles que superan la velocidad de locomoción de cualquier especie. Abstraer un problema, identificar sus principios fundamentales e independizarlo de la solución particular que se encontró para él en el reino de los seres vivos, nos ha llevado a desarrollar microscopios que ven átomos, telescopios que logran captar señales del origen mismo del universo, cohetes que vuelan a la luna, aterrizan y vuelven, y dispositivos que nos permiten comunicarnos con gente en cualquier rincón del planeta.

Primero imitar, después trascender

En el estadio actual de la IA, nos hemos alejado solo un poco del camino de la imitación. Pasamos de intentar copiar el fenómeno de

la inteligencia humana a inspirarnos en el órgano que la produce: el cerebro. Las redes neuronales, construidas emulando la estructura del cerebro humano, son la matriz sobre la que se desarrolla casi toda la IA presente. Esto puede cambiar en cualquier momento, pues ya empiezan a asomar algunas limitaciones puntuales de esta tecnología. La arquitectura de las redes actuales no se adapta bien a ciertos problemas, como la criptografía avanzada, que es una de las principales impulsoras de métodos alternativos como la computación cuántica. Este posiblemente sea el viaje transatlántico o espacial de la inteligencia. Un lugar en el que ya no sea suficiente con imitarnos, sino para el que harán falta nuevos principios, nuevas arquitecturas. Es de suponer que las redes neuronales delegarán en algún momento su reinado sobre la IA en otras formas de resolver el problema que hoy ni siquiera podemos concebir y, probablemente, las IA mismas sean quienes las descubran.

Pero las redes neuronales presentan también algunas diferencias importantes en relación con el cerebro. Su cómputo se basa en ceros y unos, en conexiones estables y duraderas, con bajísimo nivel de ruido y sin una impronta de estructura previa. El cerebro humano es lo contrario: sus neuronas tienen una dinámica compleja con todo un rango de gradaciones. Sus conexiones se hacen y deshacen permanentemente, hay un ruido de fondo que es fundamental para el buen funcionamiento cerebral y, además, el cerebro, desde el día en que nacemos, tiene una arquitectura muy definida que incide en la manera en la que vemos, en la que oímos, en la que nos movemos y en la que pensamos. Las redes neuronales empiezan en una *tabula rasa*. El cerebro humano es todo menos eso.

Hasta la llegada de las redes neuronales, la IA se resolvía programando una computadora. Un programa es una serie de instrucciones en las que se le indica paso a paso a la máquina lo que tiene que hacer. Se sabe exactamente qué hace, y cómo lo hace. En cambio, las redes neuronales son estructuras versátiles que aprenden sobre la base de grandes volúmenes de datos y encuentran soluciones a problemas que no solo no conocemos, sino que en muchos casos no podríamos visualizar. Como hemos visto en el capítulo anterior, una red neuronal descubre atributos, o representaciones intermedias, que le permiten aprender a realizar tareas. Muchas veces encuentran una forma particular de hacerlo que no se nos hubiera ocurrido jamás. Y otras tantas, estos circuitos, que se han adquirido en el entrenamiento para resolver un problema, les permiten resolver otros inesperados, sin que quien desarrolló esa red neuronal hubiese podido adivinarlo de antemano. Por eso, esta generación de IA es sorprendente. Las redes neuronales son una caja de sorpresas y eso, a la vez, les da un halo de misterio. Heredan el interés que siempre ha despertado en nosotros nuestra propia actividad cerebral. De hecho, algunas IA empiezan a estudiarse a sí mismas. Son neurocientíficas artificiales que indagan sobre sus propias representaciones para entender cómo funciona su «mente» y su «cerebro».

En las secciones anteriores, hemos esbozado lo que parece una paradoja. Conocemos muy poco del cerebro humano, no sabemos cómo son los mecanismos de la inteligencia, pero podemos ensamblar máquinas y programas que empiezan a dar muestras de poseerla. Replicar la mente humana sin entenderla del todo hace que sea complejo advertir qué principios hay detrás, qué piezas faltan y, además, deja mucho lugar a las sorpresas. Aparecen propiedades emergentes que no comprendemos del todo. Esta forma arriesgada, y quizá poco inteligente, de acercarnos a la IA la vuelve bastante impredecible.

No es la primera vez en la historia que creamos algo sin saber de que se trata o cómo funciona. En la prehistoria el hombre primitivo descubrió el fuego. Sabía encenderlo y apagarlo, y conocía los beneficios que podía obtener de él, desde cocinar alimentos hasta calentarse en los días fríos. También entendía su peligro y podía protegerse de que le hiciera daño. Sin embargo, desconocía por completo su naturaleza: qué era, por qué se comportaba de esa manera, y por qué producía esos efectos. Haber aprendido a manipular el fuego nos enseñó mucho de él. Con la inteligencia tal vez pase algo parecido, que el ejercicio de encenderla, apagarla, manipularla y usarla, aún sin terminar de comprender su naturaleza, nos lleve, finalmente, a descifrarla. De ser así, habremos honrado a Turing, que concibió este campo precisamente con ese anhelo.

Educar una máquina

¿Cómo descubre una red neuronal las palabras, las ideas, el estilo de Shakespeare o el buen uso del lunfardo? ¿Cómo se entrena una mente artificial? Esta pregunta nos remite a una vieja discusión filosófica, en el corazón de la visión socrática, sobre cómo se entrena nuestra propia mente. Un camino posible es el de AlphaGo, que se entrenó estudiando millones de partidas. Este es el camino más convencional pero no el único. Hay otra forma muy distinta para entrenar la inteligencia. La de Sócrates y Platón. Aprender preguntando, sin que nadie nos enseñe. Pensemos en este problema, por ejemplo: un ascensor tiene dos botones, uno para subir siete pisos y otro para bajar dos. Hay que ir al piso 24. ¿Cómo vamos? Encontrar la solución lleva un rato de cálculo interno, de pensar la estrategia, de identificar, entre todo nuestro conocimiento, aquel que es relevante para resolver problemas de este estilo. Y así, sin que nadie nos lo diga, llegamos a una solución. De este modo, según Sócrates, descubrimos todo lo que sabemos. Y así aprenden muchas de las inteligencias artificiales más poderosas. Solas. Sin que nadie les enseñe.

Esta es la diferencia sustancial entre AlphaGo y AlphaZero. La primera aprendió observando el comportamiento humano. La segunda aprendió sola. Jugando contra ella misma. Nosotros también utilizamos estas dos formas de aprendizaje y descubrimiento. Nos parecemos a AlphaGo cuando buscamos en Google o consultamos un libro para informarnos sobre algo que no conocemos. Otras veces, tomamos el camino de AlphaZero. Aceptamos que no entendemos algo, barajamos las cartas de nuestras ideas y logramos llegar a un resultado gracias a la capacidad de enseñarnos a nosotros mismos. Aprendemos simulando la realidad, sin necesidad de experimentarla. ¿No es fascinante pensar que sin agregar información nueva, nuestra mente, en un momento «Ajá», sepa algo que minutos antes no sabía?

Los humanos y otros animales aprendemos «por simulación» en dos lugares paradigmáticos: el juego y el sueño. En el juego infantil se descubre la coordinación sensoriomotora, pero también se aprende a calcular, a argumentar y otros cimientos de la cognición social como la provocación, la trampa, el desafío. Pintar una pared es, para un niño pequeño, una manera de experimentar con colores y materiales, pero también con la psicología de la libertad y del enfado. Algo similar ocurre durante el sueño, un espacio en el que exploramos hipótesis: ¿Qué pasa si me muero? ¿Qué pasa si alguien me deja? El juego y los sueños ponen a nuestra disposición escenarios que la vida no nos da la oportunidad de probar.

La capacidad de enseñarnos a nosotros mismos, de descubrir por el ejercicio de exploración y no de adquisición de información permite que el conocimiento vaya aumentando generación tras generación. Y esto pone en jaque a una intuición común sobre el límite de la inteligencia de las máquinas. Solemos pensar que ellas no pueden superar la inteligencia humana porque, justamente, fueron pensadas, diseñadas y entrenadas por personas. Ese razonamiento esconde una gran falacia que ya visitamos: sabemos que el discípulo puede superar al maestro. Las redes neuronales pueden identificar atributos que son indistinguibles para nosotros y así realizar tareas con habilidad sobrehumana. Pueden descubrir y aprender cosas que nadie les ha enseñado. La IA, con esta capacidad, puede llegar a lugares que sus maestros no podemos siquiera imaginar.

Un halo de misterio

Cuando un mago nos engaña con una ilusión, nos intriga saber cómo la hace. Sabemos que ha hecho un truco, pero mientras no lo descifremos, sentimos sorpresa y fascinación. En los inicios de la IA, se buscaba «entender el truco» de la inteligencia para así expresarla como una serie de instrucciones que se asemejaban a una receta de cocina. Por eso, en esos días, las IA perdían su encanto en el preciso momento en que lograban algo. A medida que la ciencia avanzaba, la delimitación entre lo que era y no era inteligencia iba cambiando, el desafío se renovaba y la definición funcionaba como una suerte de idea aspiracional: inteligencia es todo lo que las máquinas no hacen.

La camada de inteligencias artificiales que surgieron a partir del aprendizaje profundo y las inteligencias generativas alcanzaron logros e hicieron avances que seguimos sin entender del todo cómo se consiguieron, y es esa zona todavía indescifrable la que mantiene vivo el misterio y nos permite sentir que son, finalmente, inteligentes. Esta es una nueva forma de vincularnos con ellas, esta vez desde la emoción. No se trata ahora de entender la inteligencia, sino de cómo nos hace sentir. Y aquí la esencia es el misterio.

Al jugador de ajedrez cubano José Raúl Capablanca, le preguntaron en una ocasión cuántas jugadas calculaba antes de decidir qué pieza mover. Su respuesta fue: «Una sola, la mejor». Con su enigmática contestación, Capablanca abonaba el paradigma misterioso de la inteligencia: su cerebro operaba gracias a una red neuronal entrenada para determinar cuál era la mejor jugada pero no podía explicarla; ahí radicaba su halo de intriga. De la misma forma, en una de las primeras entrevistas a un jovencísimo Fernando Alonso, le preguntaron qué pensaba mientras conducía a 300 kilómetros por hora, mientras manejaba una cantidad descomunal de botones y pedales en fracciones de segundo. Su respuesta fue bastante sintética: «No pienso».

Turing representa el pensamiento consciente, el menos misterioso, el que nos permite explicar cómo hacemos las cosas. Por otra parte estarían Capablanca y Alonso, representando la dimensión inconsciente, la que interviene cuando nos enamoramos, alguien nos cae bien, y todas las cosas que sabemos o hacemos sin saber cómo o por qué. El instinto y las corazonadas están en el núcleo del pensamiento inconsciente, de las capas profundas de la red neuronal del cerebro humano. Mientras Turochamp imitaba el pensamiento consciente, las *deep learning* intentan emular el pensamiento inconsciente. Veremos cuando nos acerquemos al presente y al futuro, que una nueva generación de IA combina ya estos dos mundos, asemejándose un poco más al cerebro humano, en la medida en que logra articular ambas formas del pensamiento.

En la década de 1980, el informático austriaco Hans Moravec observó que para las máquinas es más difícil aprender algunas tareas aparentemente simples que otras mucho más complejas. Moravec

introdujo la paradoja que lleva su nombre y aborda la relación entre el pensamiento consciente e inconsciente. Dice así: «Es relativamente fácil hacer que las computadoras muestren capacidades similares a las de un humano adulto en test de inteligencia o en el juego de damas, y difícil o imposible lograr que posean las habilidades perceptivas y motrices de un bebé de un año». Según Moravec, las habilidades que resultan difíciles para los humanos, como resolver problemas matemáticos o programar, son en realidad más fáciles de realizar para las computadoras. Por otro lado, las habilidades que nos parecen simples, como caminar o agarrar un huevo con la fuerza justa para que no se caiga ni se rompa, resultan extremadamente difíciles de programar en una máquina.

Esto se debe a que el razonamiento, la planificación estratégica y la resolución de problemas abstractos son relativamente recientes en términos evolutivos, mientras que las habilidades motoras y perceptivas, como movernos, tomar objetos o reconocer caras, son mucho más antiguas; esas capacidades de nuestro pensamiento operan en una capa profunda sin que seamos conscientes de lo que está sucediendo.

Un ejemplo claro de esto se dio en la historia del ajedrez: resultó más fácil crear un programa capaz de pensar en las jugadas que uno que pudiera levantar y mover las piezas. La victoria de Deep Blue sobre Kaspárov mostró con claridad estas dos caras de la IA: una máquina decidía de manera impecable las movidas, pero era necesaria una persona que ejecutara por ella los movimientos en el tablero con la precisión y destreza de las que solo un ser humano disponía.

Sobre la efectividad y la empatía

La IA hasta hoy se utilizaba principalmente para resolver cuestiones operativas. Por ejemplo, pilotear un avión, optimizar los semáforos de una ciudad, operar con precisión un órgano humano o identificar un tumor en una imagen médica. En esas instancias, no nos importa tanto entender cómo logran lo que hacen, sino simplemente que lo hagan bien. Los programas de ajedrez actuales que

superan a los mejores humanos también nos sorprenden: no piensan como nosotros y por eso llegan a lugares que nos resultan inalcanzables. Pero este conglomerado de máquinas útiles y eficientes tiene una limitación: no son empáticas. Como el genio que nunca fracasa, que resuelve situaciones que nadie más podría resolver, genera admiración pero no cercanía. A los humanos nos gusta entendernos, jugar y lidiar con otros humanos fallidos como nosotros. Nos gustamos a nosotros mismos y nos gustan otras especies o máquinas mientras podamos proyectarnos e identificarnos con ellas.

Como ya hemos visto, con la llegada de los LLM la IA se interna en terrenos novedosos, mucho más cercanos a la creatividad y al ingenio que a las tareas mecánicas. Adquiere una forma de conversación en apariencia humana, que nos genera una inesperada sensación de cercanía. Ya había aparecido un anticipo en Eliza, con la que todos querían conversar, no porque fuese extraordinaria ni porque hiciese cálculos virtuosos, sino simplemente porque parecía sorprendentemente humana. Este es el cierre del bucle: en los experimentos de Turing y en Eliza, la IA había sido un intento por entender lo más sorprendente de la condición humana. Luego la investigación en IA se fue durante décadas de excursión a un mundo pragmático y eficiente donde lo relevante era resolver bien un problema, sin importar la manera. De repente, el camino nos trae de vuelta a casa, al lugar de la conversación, del juego de imitación, de una máquina que se confunde con uno de nosotros. En nuestros aciertos pero también en nuestras imprecisiones.

Nos vamos acostumbrando a vincularnos con las inteligencias artificiales porque ya hablan con nosotros, hacen resúmenes, dan consejos y juegan. Apreciamos la respuesta que nos ofrece ChatGPT porque empatiza con nuestra forma de escribir y de percibir la escritura. Sucede así porque estas IA se han entrenado con datos de la cultura humana, que han digerido entera, y las cosas que se les ocurren se estructuran sobre todo ese conocimiento.

Sin embargo, sabemos que la imitación está muy cerca de la impostura, y detalles ínfimos conducen del amor al desprecio. Si no, que le pregunten a una rata por qué leves diferencias en su apariencia la hacen repulsiva frente a una ardilla, que la mayoría de

gente encuentra adorable. Y es que la curiosidad que nos genera vincularnos con otras inteligencias (o simplemente con otros entes) se encuentra con roces y reparos bastante estereotipados. En 1970, el robotista japonés Masahiro Mori llamó «valle inquietante» a la respuesta emocional negativa que experimenta una persona cuando se encuentra con un objeto o un ser humanoide que es casi, pero no del todo, realista. A medida que los robots humanoides se vuelven más parecidos a los seres humanos en apariencia y comportamiento, generalmente despiertan una mayor empatía y aceptación por parte de las personas. Sin embargo, hay un punto en el que la semejanza se acerca lo suficiente a la realidad como para resultar familiar, pero con algunos detalles o características sutiles que delatan la casi perfecta impostura. Justo en ese punto, sentimos una sensación de inquietud y repulsión hacia el robot, y la empatía se desploma. No hay nada más molesto que algo que se parece mucho a una persona, sin serlo. Lo ligeramente falso suele generar mucho malestar.

El «valle inquietante» de Mori se mide con la vara de la imitación perfecta. ¿Y si un robot pasase del otro lado de esa vara? ¿Puede la imitación de una inteligencia ser más inteligente, e incluso de aspecto más humano, que los humanos que la han creado? Empezamos esta sección viendo que hemos fabricado microscopios y telescopios que nos permiten ver lo que el ojo no ve, y máquinas que vuelan más alto y más lejos que cualquier ave. ¿Qué va a pasar con la inteligencia artificial cuando llegue a sitios que la nuestra es incapaz de alcanzar y quizá hasta de concebir? En este caso parece haber algo sustancialmente distinto que en el resto de máquinas, autómatas y artefactos. Algo que es transversal a todo el contenido de este libro y que aparece en cada capítulo en distintas manifestaciones. La inteligencia es el rasgo más preciado que tenemos, el orgullo de nuestra especie. Si mandáramos al espacio un arca con creaciones humanas, no mandaríamos nuestro copioso sudor como forma de mantener la regulación térmica, ni disecciones de rodillas y codos para mostrar la versatilidad articulatoria de un miembro. Irían canciones, poemas, cartas, pinturas. En fin, distintas expresiones de la inteligencia y la cultura.

Por esto mismo a nadie le ofende que un auto sea más rápido que nosotros, pero sí nos inquieta que una máquina piense mejor. Porque nos toca en la fibra más íntima. ¿Qué sentiríamos, en definitiva, si hubiese una especie de entes artificiales mucho más inteligentes que nosotros? Parte del temor es evidente. Serán mejores en aquella cualidad que nos hizo lo que somos. Porque, para bien y para mal, edificios, catedrales, imprentas, basureros, guerras, bombas, cartas de amor, telescopios, teoremas y circos son ejemplos de nuestras creaciones. No somos una especie particularmente rápida, ni fuerte, ni resiliente. La inteligencia es la herramienta con la que hemos creído gobernar el mundo e impuesto nuestra voluntad sobre las demás criaturas. Hacemos bien en temer lo que pueda pasar cuando alguien o algo nos supere, y pueda ser quien decida si andamos sueltos, con correas, o enjaulados.

Ese oscuro objeto del deseo

La función de valor, que ya presentamos hace poco, es el eslabón fundamental del mecanismo de aprendizaje que está en el corazón de la IA. La regla es simple: hay que lograr optimizar algo. Esto puede ser ganar al ajedrez, lograr que la pelota que lanzamos caiga justo en el aro, que un video en TikTok sea visto por mucha gente o ganar una elección parlamentaria; da igual. El punto es que ejecutamos acciones y vemos los resultados. Si agregamos, por ejemplo, texto a un video y nos damos cuenta de que esto hace que lo vea más gente, repetiremos el proceso. Esto tan simple, repetido en millones y millones de ensayos, permite encontrar la receta justa que hace que un movimiento del brazo sea el preciso para que la pelota entre en el aro o que mover un caballo de una manera imprevista sea la jugada ganadora. El método funciona por tres razones vitales que conviene tener en mente, porque no estarán siempre presentes y ahí el plan de aprendizaje se complica:

1. Poder observar de manera clara los resultados de la acción que llevamos a cabo.

2. Disponer de una función de valor que nos permita medir de manera precisa si ese acto ha generado efectos positivos o negativos, para fortalecer o debilitar los parámetros del modelo.
3. Disponer de una cantidad gigantesca de estas pruebas y errores, para llegar a los valores ideales para esos miles de parámetros que permiten predecir cómo es conveniente seguir.

El asunto más espinoso de los tres suele ser el segundo: definir cuál es la función de valor que establece qué es lo que queremos lograr. Y esto es porque la función de valoración es arbitraria, define un objetivo y en cierta manera una moral, o una teleología. Condensa la base práctica y filosófica por la que, en última instancia, hacemos las cosas.

En cuanto nos apartamos de dominios claramente delimitados, objetivamente medibles, se puede volver muy borroso establecer cuál es la función de valor que hay que optimizar. ¿Cómo conducir un auto? ¿Cómo cuidar a una persona mayor? ¿Qué comer para sentirnos mejor? En cuanto queremos que una IA nos dé respuestas sobre cuestiones que cambian nuestra vida, este asunto adopta otra relevancia: ¿cuál es la función de valor de la vida? El dicho popular «cuidado con lo que deseamos» refleja bien un riesgo relacionado con esto, tanto en lo humano como en los artificios que hemos inventado. Es el peligro de la función de valor equivocada: perseguir una meta y darnos cuenta al alcanzarla de que no era lo que esperábamos; o detectar que, en el proceso, hemos sacrificado otros aspectos de la vida que eran, a fin de cuentas, mucho más relevantes. Los ejemplos son de lo más variado: con quién pasamos nuestro tiempo, las cosas por las que nos preocupamos, en qué invertimos nuestros ingresos, lo que estudiamos o no, cómo reaccionamos frente a una discusión callejera. En cada uno de estos ejemplos, el cerebro decide, sin previo aviso, cuál es la función de valor que va a optimizar. Pasamos más tiempo pensando cómo alcanzar un objetivo que nos hemos propuesto que preguntándonos si tal vez no deberíamos cambiar la meta que perseguimos.

El informático británico Stuart Russell se ha ocupado de este asunto identificando cuáles son las ideas para que una IA esté verdaderamente alineada con los objetivos, más grandes y trascendentes, de la especie humana. Para eso, ha tomado el siguiente ejemplo: «Si introduces un objetivo en una máquina, algo simple como "traer el café", la máquina se dice a sí misma: "Bueno, ¿cómo podría fallar en traer el café? Alguien podría apagarme. De acuerdo, tengo que tomar medidas para evitar eso. Desactivaré mi interruptor de apagado. Haré cualquier cosa para defenderme de lo que interfiera con mi objetivo". Esta búsqueda obstinada de un modo muy defensivo para cumplir un objetivo no está alineada con los principios de la especie humana». Este es el problema de una función de valor equivocada que establece un objetivo simple (llegar a tiempo a un destino, llevar un café) olvidando que hay otro conjunto de principios con los que este objetivo tiene que convivir: no atropellar a nadie por el camino, no hacer daño, no matar... Algunos de estos principios son obvios. Otros no tanto, y esa es la razón por la que no logramos ponernos de acuerdo acerca de ellos desde hace miles de años. Mientras, Russell señala tres principios que pueden guiarnos en este lío: el altruismo (el gran objetivo de la máquina tiene que apuntar a maximizar la realización de los valores humanos), la humildad (la máquina inicialmente siente incertidumbre acerca de cuáles son las preferencias humanas) y el hecho de que el aprendizaje debe provenir de los humanos (que la máquina evite otras fuentes de información que no sean humanas).

El ejemplo del café pone de manifiesto un problema que está en la esencia de la filosofía, la teleología y el derecho: la ambigüedad en las interpretaciones y las indefectibles omisiones en cualquier texto que pretenda sintetizar los fundamentos de la moral. Para mostrar que esto es algo esencial al lenguaje, el profesor de computación científica de la Universidad de Harvard, David Malan, pone en evidencia el problema de la imprecisión y la ambigüedad del lenguaje en un dominio muchísimo más simple que el de la moral. ¿Cómo dar las indicaciones para que otra persona haga un sándwich de pan y mermelada? En un experimento que hizo con

sus alumnos, les pidió que le dieran las instrucciones paso a paso para hacerlo. David Malan cumple las instrucciones a rajatabla (como un robot, diríamos), y el resultado es que una y otra vez fracasa en un objetivo tan simple. Después de muchos intentos fallidos y panes desperdiciados, el experimento mostró que para que la comunicación funcione bien, aun en las tareas más simples, una persona o una inteligencia artificial tienen que estar embebidas de todo ese contexto que muchas veces se da por sentado. Por ejemplo, si uno de sus estudiantes dictaba como primer paso «Abra la bolsa de pan» se asume que el profesor no tenía que romper o dejar caer las rebanadas en el proceso, pero esto no se expresaba en la instrucción.

Un niño posee numerosas intuiciones sobre cómo realizar tareas, y es probable que, al preparar un sándwich, evite muchos de los malentendidos en los que una IA podría caer. Afinar y precisar el proceso de instrucción resulta clave en este momento en el que las IA resuelven cosas mucho más relevantes que hacer un pan con mermelada. Pequeños errores u omisiones pueden llevar a problemas sustanciales.

Lo que estamos viendo parece un inconveniente de las máquinas, de las inteligencias artificiales. Pero, en realidad, es un problema en la esencia de lo humano que solo se hace explícito cuando aparecen programas que realizan acciones que ponderan nuestros valores. Es que solemos pensar en la IA como una suerte de alienígena que viene de fuera a rivalizar con nuestra especie. Esta interpretación pasa por alto que las máquinas fueron hechas por nosotros y aprenden de nuestros textos y acciones. Con cada uno de ellos van heredando nuestros principios. Incluso cuando aprenden solas, son humanos quienes escriben la función de valor, al menos por ahora. Nuestra impronta está tan presente que, en algún punto, nos obliga a preguntarnos cuáles son nuestros valores y hasta qué punto son convencionales o universales. Lejos de ser alienígena, la IA funciona como un espejo que refleja todas las maravillosas capacidades humanas que han sido necesarias para llegar a esos desarrollos, pero también los defectos y vicios de nuestra propia condición.