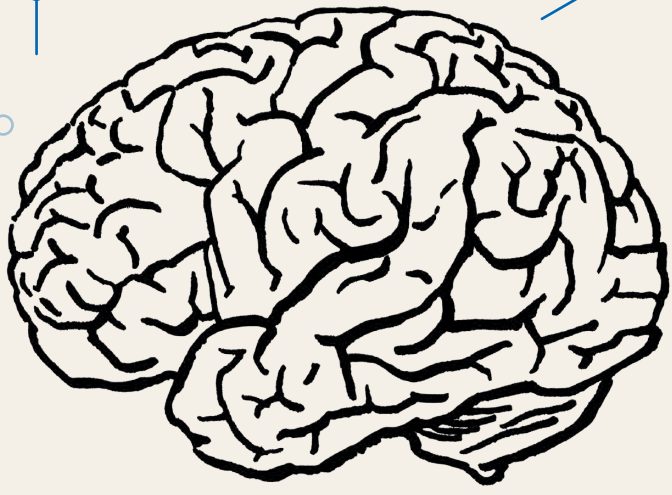
METATEMAS

Jeff Hawkins

Mil cerebros

Una nueva teoría de la inteligencia

PRÓLOGO DE RICHARD DAWKINS



TUSQUETS EDITORES

Jeff Hawkins

MIL CEREBROS

Una nueva teoría de la inteligencia

Prólogo de Richard Dawkins

Traducción de Ambrosio García Leal

TUSQUETS
EDITORES

Título original: *A Thousand Brains*

1.ª edición: marzo de 2023

Tusquets Editores, S.A. – Avda. Diagonal, 662-664 – 08034 Barcelona
www.tusquetseditores.com
ISBN: 978-84-1107-249-6
Depósito legal: B. 1.726-2023
Fotocomposición: David Pablo
Impresión y encuadernación: CPI Black Print
Impreso en España

Índice

Primera parte
Una nueva comprensión del cerebro

Las células que tienes en la cabeza están leyendo estas palabras. Piensa en lo extraordinario que es eso. Las células son simples. Una sola célula no puede leer, ni pensar ni hacer casi nada. Sin embargo, si juntamos suficientes células para formar un cerebro, no solo leen libros, sino que los escriben. Diseñan edificios, inventan tecnologías y descifran los misterios del universo. Cómo un cerebro hecho de células simples puede generar inteligencia es una cuestión de profundo interés, y sigue siendo un misterio.

Comprender el funcionamiento del cerebro se considera uno de los grandes desafíos de la humanidad. Este propósito ha generado decenas de iniciativas nacionales e internacionales, como el Proyecto Cerebro Humano europeo (HBP, por sus siglas en inglés) o la Iniciativa Internacional del Cerebro (IBI, por sus siglas en inglés). Decenas de miles de neurólogos trabajan en decenas de especialidades, en prácticamente todos los países del mundo, tratando de comprender el cerebro. Aunque los neurólogos estudian los cerebros de diferentes animales y plantean preguntas variadas, el objetivo final de la neurociencia es aprender cómo el cerebro humano da origen a la inteligencia humana.

Mi afirmación de que el cerebro humano sigue siendo un misterio puede que sorprenda a más de uno. Cada año se anuncian nuevos descubrimientos relacionados con el cerebro, se publican nuevos libros sobre el tema y los investigadores de campos relacionados, como la inteligencia artificial, afirman que sus

creaciones se acercan a la inteligencia de, por ejemplo, un ratón o un gato. Sería fácil concluir de esto que los científicos tienen una idea bastante buena de cómo funciona el cerebro. Pero si preguntamos a los neurólogos, casi todos admitirían que todavía estamos en la oscuridad. Hemos reunido una gran cantidad de conocimientos y hechos sobre el cerebro, pero apenas comprendemos su funcionamiento.

En 1979, Francis Crick, famoso por su trabajo sobre el ADN, escribió un artículo sobre el estado de la neurociencia titulado «Reflexiones en torno al cerebro». En él describía la gran cantidad de datos que los científicos habían recopilado sobre el cerebro; sin embargo, «a pesar de la constante acumulación de conocimientos detallados, el funcionamiento del cerebro humano sigue siendo insondablemente misterioso». Y añadía: «Lo que llama la atención es la ausencia de un marco amplio de ideas para interpretar estos resultados».

Crick remarcaba que los científicos habían estado recopilando datos sobre el cerebro durante décadas. Sabían muchísimo del cerebro. Pero nadie había dado con la manera de ensamblar todo ese conocimiento en algo significativo. El cerebro era como un rompecabezas gigante con miles de piezas. Teníamos las piezas delante de nosotros, pero no podíamos darles sentido. Nadie sabía por dónde debería ir la solución. Según Crick, el cerebro era un misterio no porque no hubiéramos recopilado suficiente información, sino porque no sabíamos cómo organizar las piezas que ya teníamos. En los cuarenta años transcurridos desde que Crick escribió su artículo se han hecho muchos descubrimientos relevantes sobre el cerebro, de algunos de los cuales hablaré más adelante, pero lo que allí decía sigue siendo cierto en términos generales. Cómo surge la inteligencia de las células de nuestra cabeza sigue siendo un misterio insondable. A medida que se reúnen más piezas del rompecabezas cada año, a veces parece que nos estuviéramos alejando de la comprensión del cerebro, en lugar de acercarnos.

Leí el artículo de Crick cuando era joven, y fue inspirador. Sentí que podríamos resolver el misterio del cerebro a lo largo de mi vida, y he perseguido ese objetivo desde entonces. En los últimos diez años he dirigido un equipo de investigación en Silicon Valley que estudia una parte del cerebro llamada neocórtex. El neocórtex abarca alrededor del 70 por ciento del volumen del cerebro humano, y es responsable de todo lo que asociamos con la inteligencia, desde nuestros sentidos de la vista, el tacto y el oído hasta el lenguaje en todas sus formas, incluyendo el pensamiento abstracto, como las matemáticas y la filosofía. El objetivo de nuestra investigación es comprender cómo funciona el neocórtex con suficiente detalle para poder explicar la biología del cerebro y construir máquinas inteligentes que se basen en los mismos principios.

A principios de 2016, el progreso de nuestra investigación se aceleró drásticamente. Tuvimos un gran avance en nuestra comprensión del cerebro. Nos dimos cuenta de que habíamos pasado por alto un ingrediente clave. Con esta nueva perspectiva, pudimos ver cómo encajaban las piezas del rompecabezas. En otras palabras, creo que descubrimos el marco al que se refería Crick, un marco que no solo explique las bases del funcionamiento del neocórtex, sino que también propicie una nueva concepción de la inteligencia. Todavía no tenemos una teoría completa del cerebro, ni mucho menos. Los campos científicos suelen partir de un marco teórico, y solo más tarde se resuelven los detalles. La teoría de la evolución de Darwin puede que sea el ejemplo más famoso. Darwin propuso una nueva y audaz concepción del origen de las especies, pero los detalles, como el funcionamiento de los genes y el ADN, no se conocerían hasta muchos años después.

Para ser inteligente, el cerebro tiene que aprender muchas cosas sobre el mundo. No me refiero solo a lo que aprendemos en la escuela, sino a cosas básicas, como el aspecto, el sonido y el tacto de objetos cotidianos. Tenemos que aprender cómo se com-

portan esos objetos, desde cómo se abre y cierra una puerta hasta qué hacen las aplicaciones de nuestros teléfonos inteligentes cuando tocamos la pantalla. Necesitamos saber dónde se encuentra cada cosa en el mundo, desde dónde guardamos nuestras pertenencias personales en casa hasta dónde están la biblioteca y la oficina de correos de nuestra ciudad. Y, por supuesto, aprendemos el significado de conceptos de nivel superior, como «compasión» o «gobierno». Además de todo esto, cada uno de nosotros aprende el significado de decenas de miles de palabras. Cada uno de nosotros posee una tremenda cantidad de conocimiento del mundo. Algunas de nuestras habilidades básicas vienen determinadas por nuestros genes, como comer o retroceder ante el dolor. Pero la mayor parte de lo que sabemos sobre el mundo se aprende.

Los científicos dicen que el cerebro aprende un modelo del mundo. La palabra «modelo» implica que lo que sabemos no se almacena simplemente como un montón de hechos, sino que se organiza de manera que refleja la estructura del mundo y todo lo que contiene. Por ejemplo, para saber qué es una bicicleta no recordamos una lista de datos sobre bicicletas. En vez de eso, nuestro cerebro crea un modelo de bicicleta que incluye las diferentes partes, cómo están dispuestas las partes entre sí y cómo se mueven y trabajan juntas. Para reconocer algo, primero debemos aprender cómo se ve y cómo se siente, y para lograr objetivos, debemos aprender cómo se comportan las cosas en el mundo cuando interactuamos con ellas. La inteligencia está íntimamente ligada al modelo cerebral del mundo; por lo tanto, para entender cómo crea inteligencia, tenemos que averiguar cómo el cerebro, hecho de células simples, aprende un modelo del mundo y todo lo que hay en él.

Nuestro descubrimiento de 2016 explica cómo aprende el cerebro este modelo. Dedujimos que el neocórtex almacena todo lo que sabemos, todo nuestro conocimiento, valiéndose de lo que llamamos marcos de referencia. Explicaré esto con más detalle

más adelante, pero por ahora consideremos un mapa de papel como una analogía. Un mapa es un tipo de modelo: un mapa de un pueblo es un modelo del pueblo, y las líneas de cuadrícula, como las de latitud y longitud, son un tipo de marco de referencia. Las líneas de cuadrícula de un mapa, su marco de referencia, proporcionan la estructura del mapa. Un marco de referencia nos dice dónde se sitúan unas cosas respecto de otras, y puede decirnos cómo lograr objetivos, como por ejemplo ir de un sitio a otro. Vimos que el modelo cerebral del mundo se construye empleando marcos de referencia similares a mapas. No un marco de referencia, sino cientos de miles de ellos. De hecho, ahora comprendemos que la mayoría de las células del neocórtex se dedica a crear y manipular marcos de referencia, de los que el cerebro se vale para planificar y pensar.

Con esta nueva perspectiva, comenzaron a vislumbrarse las respuestas a algunos de los interrogantes más importantes de la neurociencia, como de qué manera nuestras diversas entradas sensoriales confluyen en una experiencia singular, o qué ocurre cuando pensamos, o cómo pueden dos personas llegar a diferentes creencias a partir de las mismas observaciones, o por qué tenemos un sentido del yo.

Este libro cuenta la historia de estos descubrimientos y las implicaciones que tienen para nuestro futuro. La mayor parte del material se ha publicado en revistas científicas. Proporciono enlaces a estos documentos al final del libro. Sin embargo, los artículos científicos no son adecuados para explicar teorías a gran escala, especialmente de una manera que los no especialistas puedan entender.

He dividido el libro en tres partes. En la primera describo nuestra teoría de los marcos de referencia, la que hemos llamado teoría de los mil cerebros. Esta teoría se basa en parte en la deducción lógica, así que detallaré los pasos que seguimos para llegar a nuestras conclusiones. También ofreceré algunos antece-

dentes históricos que ayudan a ver cómo se relaciona la teoría con la historia del pensamiento sobre el cerebro. Al final de esta primera parte, espero que entiendas lo que está pasando en tu cabeza cuando piensas y actúas en el mundo, y lo que significa ser inteligente.

La segunda parte del libro trata de la inteligencia artificial. El siglo XXI se verá transformado por las máquinas inteligentes de la misma manera que el siglo XX lo fue por las computadoras. La teoría de los mil cerebros explica por qué la IA actual aún no es inteligente, y qué debemos hacer para crear máquinas inteligentes de verdad. Describo cómo serán las máquinas inteligentes del futuro y cómo podríamos usarlas. También explico por qué algunas máquinas serán conscientes y qué debemos hacer al respecto, si es que debemos hacer algo. Finalmente, a mucha gente le preocupa que las máquinas inteligentes sean un riesgo existencial, que estemos a punto de crear una tecnología que destruirá a la humanidad. No estoy de acuerdo. Nuestros descubrimientos ilustran por qué la inteligencia artificial, en sí misma, es benigna. Pero, como tecnología poderosa que es, el riesgo radica en los usos que podrían hacerse de ella.

En la tercera parte del libro analizo la condición humana desde la perspectiva del cerebro y la inteligencia. El modelo cerebral del mundo incluye un modelo de nosotros mismos. Esto conduce a la extraña constatación de que lo que tú y yo percibimos, momento a momento, es una simulación del mundo, no el mundo real. Una consecuencia de la teoría de los mil cerebros es que nuestras creencias sobre el mundo pueden ser falsas. Explico cómo puede ocurrir esto, por qué las creencias falsas pueden ser difíciles de erradicar, y cómo dichas creencias falsas combinadas con nuestras emociones más primitivas son una amenaza para nuestra supervivencia a largo plazo.

Los capítulos finales tratan de lo que considero que es la elección más importante que afrontaremos como especie. Hay

dos formas de pensar en nosotros mismos. Una es como organismos biológicos, producto de la evolución y la selección natural. Desde este punto de vista, los seres humanos venimos definidos por nuestros genes, y el propósito de nuestra vida es replicarlos. Pero ahora estamos saliendo de nuestro pasado puramente biológico. Nos hemos convertido en una especie inteligente. Somos la primera especie de la Tierra en conocer el tamaño y la edad del universo. Somos la primera especie en saber cómo evolucionó la Tierra y cómo llegamos a existir. Somos la primera especie en desarrollar herramientas que nos permiten explorar el universo y conocer sus secretos. Desde este punto de vista, los seres humanos estamos definidos por nuestra inteligencia y nuestro conocimiento, no por nuestros genes. La elección a la que nos enfrentamos cuando pensamos en el futuro es si seguimos dejándonos impulsar por nuestro pasado biológico o si, por el contrario, abrazamos nuestra recién nacida inteligencia.

Es posible que no podamos hacer ambas cosas. Estamos creando tecnologías poderosas que pueden alterar fundamentalmente nuestro planeta, manipular la biología y, pronto, crear máquinas más inteligentes que nosotros. Pero aún poseemos los comportamientos primitivos que nos han llevado hasta aquí. Esta combinación es el auténtico riesgo existencial que debemos abordar. Si estamos dispuestos a aceptar la inteligencia y el conocimiento como lo que nos define, más que nuestros genes, entonces quizás podamos construir un futuro más duradero y con un propósito más noble.

El viaje que condujo a la teoría de los mil cerebros ha sido largo y tortuoso. Estudié ingeniería electrónica en la universidad, y había empezado a trabajar en Intel cuando leí el libro de Francis Crick. Tuvo un efecto tan profundo en mí que decidí cambiar de carrera y dedicar mi vida a estudiar el cerebro. Tras un intento infructuoso de conseguir un puesto que me permitiera estudiar el cerebro en Intel, solicité una plaza de posgraduado en el labora-

torio de inteligencia artificial del MIT. (Me parecía que la mejor manera de construir máquinas inteligentes era estudiar primero el cerebro.) En mis entrevistas con el profesorado del MIT, mi propuesta de crear máquinas inteligentes basadas en el cerebro fue rechazada. Me dijeron que el cerebro no era más que un ordenador embrollado que no tenía sentido estudiar. Cabizbajo, pero sin inmutarme, me inscribí en el programa de doctorado en neurociencia de la Universidad de California en Berkeley. Comencé mis estudios en enero de 1986.

Al llegar a Berkeley, pedí consejo al responsable del grupo de posgrado de neurobiología, el Dr. Frank Werblin. Me dijo que escribiera una descripción de la investigación que quería realizar para mi tesis doctoral. En el documento expliqué que quería trabajar en una teoría del neocórtex. Sabía que quería abordar el problema estudiando cómo el neocórtex hace predicciones. El profesor Werblin hizo que varios miembros de la facultad leyeran mi proyecto, y fue bien recibido. Me dijo que mis ambiciones eran admirables, que mi enfoque era sólido y que el problema en el que quería trabajar era uno de los más importantes de la ciencia, pero —y esto no me lo esperaba— no veía cómo podría perseguir mi sueño en aquel momento. Como estudiante de posgrado en neurociencia, tendría que trabajar para un profesor, haciendo un trabajo similar al suyo. Y nadie en Berkeley, ni en ningún otro lugar que conociera, estaba haciendo algo lo bastante parecido a lo que yo quería hacer.

Intentar desarrollar una teoría general del funcionamiento del cerebro se consideraba un proyecto demasiado ambicioso y, por ende, demasiado arriesgado. Si un estudiante trabajaba en esto durante cinco años y no avanzaba, podría no graduarse. También era arriesgado para los profesores, que se jugaban la titularidad. Los organismos que financian la investigación también lo consideraban demasiado arriesgado. Las propuestas de investigación centradas en la teoría se rechazaban de manera sistemática.

Podía haber trabajado en un laboratorio experimental, pero tras unas cuantas entrevistas supe que no era una buena opción para mí. Pasaría la mayor parte de mi jornada adiestrando animales, construyendo equipos experimentales y recogiendo datos. Cualquier teoría que desarrollara se limitaría a la parte del cerebro a la que se dedicaba ese laboratorio.

Durante los dos años siguientes me pasé los días en las bibliotecas universitarias leyendo un artículo de neurociencia tras otro. Leí cientos de ellos, incluyendo los artículos más importantes publicados en los últimos cincuenta años. También leí lo que pensaban psicólogos, lingüistas, matemáticos y filósofos sobre el cerebro y la inteligencia. Recibí una educación de primera clase, aunque poco convencional. Después de dos años de formación autodidacta, se hacía necesario un cambio. Se me ocurrió un plan. Volvería a trabajar en la industria durante cuatro años y luego reevaluaría mis oportunidades en el mundo académico. Así que volví a trabajar en el campo de los ordenadores personales en Silicon Valley.

Empecé a tener éxito como emprendedor. De 1988 a 1992 creé uno de los primeros ordenadores tablet, el GridPad. En 1992 fundé Palm Computing, y durante los siguientes diez años diseñé algunos de los primeros ordenadores portátiles y teléfonos inteligentes, como el PalmPilot y el Treo. Todos los que trabajaron conmigo en Palm sabían que mi corazón estaba en la neurociencia, que veía mi trabajo en la informática móvil como algo temporal. Diseñar algunos de los primeros portátiles y teléfonos inteligentes fue un trabajo apasionante. Sabía que miles de millones de personas iban a depender de esos dispositivos, pero sentía que entender el cerebro era aún más importante. Creía que la teoría del cerebro tendría un impacto positivo mayor en el futuro de la humanidad que la informática. Tenía que volver a la investigación del cerebro.

Ningún momento era bueno para marcharse, así que elegí una fecha y me fui de las empresas que ayudé a crear. Con la asistencia y el empuje de algunos amigos neurocientíficos (espe-

cialmente Bob Knight, de Berkeley; Bruno Olshausen, de la Universidad de California en Davis, y Steve Zornetzer, del Centro de Investigaciones Ames de la NASA), creé el Instituto Redwood de Neurociencia (RNI, por sus siglas en inglés) en 2002. El RNI se centraba exclusivamente en la teoría neocortical, y contaba con diez científicos a tiempo completo. Todos estábamos interesados en las teorías a gran escala del cerebro, y el RNI era uno de los pocos lugares del mundo donde este enfoque no solo se toleraba, sino que se esperaba. A lo largo de los tres años que estuve al frente del RNI, tuvimos más de un centenar de expertos visitantes, algunos de los cuales se quedaron durante días o semanas. Teníamos conferencias semanales, abiertas al público, que solían convertirse en horas de discusión y debate.

Todos los que trabajaban en el RNI, yo incluido, pensaban que era genial. Pude conocer personalmente y pasar tiempo con muchos de los principales neurocientíficos del mundo. Esto me permitió familiarizarme con múltiples campos de la neurociencia, cosa difícil de conseguir en un puesto académico típico. El problema era que yo quería dar respuesta a una serie de preguntas concretas, y no veía que se progresara hacia un consenso sobre esas cuestiones. Así que, después de tres años dirigiendo un instituto de investigación, decidí que la mejor manera de alcanzar mis objetivos era dirigir mi propio equipo de investigación.

Dado que el RNI funcionaba bien en todos los demás aspectos, decidimos trasladarlo a Berkeley. Sí, la misma institución que me dijo que no podía estudiar la teoría del cerebro decidió, diecinueve años después, que un centro de teoría del cerebro era justo lo que necesitaba. El RNI continúa hoy con otro nombre, el RCTN (Centro Redwood de Neurociencia Teórica).

Tras el traslado del RNI a Berkeley, varios colegas y yo creamos Numenta, una empresa de investigación independiente. Nuestro objetivo principal es desarrollar una teoría sobre el funcionamiento del neocórtex. Nuestro objetivo secundario es aplicar lo

que sabemos del cerebro a las máquinas que aprenden y a la inteligencia artificial. Numenta es similar a un laboratorio de investigación universitario típico, pero más flexible. Me permite dirigir un equipo, asegurarme de que todos estamos centrados en la misma tarea y probar nuevas ideas siempre que sea necesario.

Mientras escribo esto, Numenta tiene ya más de quince años, pero en algunos aspectos seguimos estando en los inicios. Tratar de averiguar cómo funciona el neocórtex es un enorme reto. Para progresar, necesitamos la flexibilidad y la concentración de un entorno de empresa emergente. También necesitamos mucha paciencia, lo cual no es típico de una empresa emergente. Nuestro primer descubrimiento significativo —cómo las neuronas hacen predicciones— se produjo en 2010, cinco años después de nuestra puesta en marcha. El descubrimiento de los marcos de referencia en el neocórtex se produjo seis años más tarde, en 2016.

En 2019 empezamos a trabajar en nuestra segunda misión, aplicar los principios del cerebro al aprendizaje automático. En ese mismo año empecé a escribir este libro, para compartir lo que hemos aprendido.

Me parece asombroso que la única cosa en el universo que sabe de la existencia del universo sea el kilo y medio de células que flota en nuestras cabezas. Me recuerda el viejo enigma: si un árbol cae en el bosque y nadie está allí para oírlo, ¿hizo algún ruido? Igualmente podemos preguntarnos: si algún universo surgió y dejó de existir sin que hubiera cerebros que supieran de su existencia, ¿existió realmente? ¿Quién lo sabría? El caso es que unos cuantos miles de millones de células suspendidas en tu cráneo saben no solo que el universo existe, sino que es inmenso y antiguo. Esas células han concebido un modelo del mundo, un conocimiento que, por lo que sabemos, no existe en ninguna otra parte. Llevo toda una vida intentando comprender cómo lo consigue el cerebro, y me apasiona lo que hemos aprendido. Espero que mi entusiasmo sea compartido. Empecemos.