La inteligencia artificial explicada a los humanos

La inteligencia artificial explicada a los humanos

Jordi Torres

Prólogo de Mateo Valero

Plataforma Editorial

Primera edición en esta colección: septiembre de 2023

Plataforma Editorial
c/ Muntaner, 269, entlo. 1ª – 08021 Barcelona
Tel.: (+34) 93 494 79 99
www.plataformaeditorial.com
info@plataformaeditorial.com

Depósito legal: B-15480-2023
ISBN: 978-84-19655-56-1
IBIC: PDZ
Printed in Spain – Impreso en España

Diseño de cubierta:
Sara Miguelena

Fotocomposición:
Grafime

El papel que se ha utilizado para imprimir este libro proviene
de explotaciones forestales controladas, donde se respetan
los valores ecológicos, sociales y el desarrollo sostenible del bosque.

Impresión:
Sagrafic

A mi familia

Índice

Prefacio

A finales de 2022 hubo un punto de inflexión en nuestra relación con la inteligencia artificial (IA) debido, en gran medida, a la aparición de diferentes programas informáticos al alcance de todos los usuarios. Estas IA permiten —a cualquier persona con acceso a Internet— generar textos e imágenes que en muchos casos es muy difícil saber si han sido creados por una IA o por un humano.

Esto avivó un interesante debate público: hacia dónde se dirige la IA y qué consecuencias puede acabar teniendo para la humanidad. Los medios de comunicación se han hecho eco de cómo la IA forma parte de prácticamente todos los aspectos de nuestra vida y de que cambiará el mundo de forma irreversible. Pese a que como sociedad estamos asimilándolo, aún no hay consenso sobre dónde nos llevará la revolución de la IA en la que nos encontramos inmersos.

Las opiniones están muy polarizadas —como sucede últimamente en casi todo—. Por un lado, hay quien cree que la IA es una aliada que podrá aportar soluciones a los grandes retos que se le presentan a nuestra sociedad. Por otro lado, están los que piensan que la IA es una enemiga de la huma-

nidad, quizás por la influencia que ha tenido la ciencia ficción y las distopías con sus máquinas con superinteligencia, generalmente antropomórficas, capaces de superar y rebelarse contra el ser humano.

En cualquier caso, toda herramienta poderosa puede ser beneficiosa o perjudicial, dependiendo de quién la utilice y con qué fines. Es decir, aunque la inteligencia artificial tiene un gran potencial para mejorar nuestra vida, su uso imprudente puede ser dañino y tener un impacto negativo en la humanidad.

En líneas generales, reinan la inquietud y la confusión entre la población. Veámoslo con un ejemplo sencillo: muchas personas, cuando navegan por Internet o abren aplicaciones en su móvil cada día, no son conscientes de que están usando una IA y que esta condiciona sus acciones. Este desconocimiento los deja totalmente a su merced.

Otro ejemplo es la aparición del chatbot, un servicio de IA gratuito con una capacidad de escritura tan sofisticada que el texto que produce es inquietantemente verosímil, tanto que parece escrito por un humano. La llegada del ChatGPT, por mencionar el más conocido, ha puesto en jaque el modo en que hemos enseñado y evaluado durante decenios a nuestros estudiantes.

El desconocimiento genera confusión, temor, rechazo. Una de las principales causas de este desconcierto (entre el público en general) es que se tiende a utilizar un lenguaje demasiado técnico cuando quien lo explica es un experto en la materia. Sin embargo, cuando quien lo explica es un

divulgador, se enfrenta al desafío de transmitir la esencia y perspectiva del tema en un lenguaje claro, lo cual puede resultar difícil, si no imposible. Y esto, sin duda, genera desasosiego y, demasiado a menudo, claudicación.

La IA ha revolucionado —y lo hará mucho más— la forma en que interactuamos con el mundo, por lo que su comprensión resulta necesaria para entender cómo funcionan las relaciones y la sociedad. Con el objetivo de explicar de manera rigurosa pero accesible el funcionamiento y el impacto de la IA, he decidido escribir un libro dirigido al público no especializado.

Como veremos, a pesar de que el progreso de la IA se debe a tres vectores clave (los avances algorítmicos, la disponibilidad de grandes cantidades de datos y la capacidad de computación), ha sido este último el que ha marcado el ritmo de crecimiento de la IA. Por ello, investigar en supercomputación resulta también de gran ayuda para poder explicar cómo ha evolucionado y cómo puede continuar evolucionando la IA.

La presente propuesta espera ser, precisamente, una explicación rigurosa y accesible para un lector sin conocimientos técnicos y científicos previos. Para ello recurro a generalizar los conceptos —sin faltar al rigor científico— con el objetivo de que el lector no experto en IA encuentre la lectura amena y comprensible. Se apuesta por un lenguaje llano para la descripción de los conceptos fundamentales a la vez que construye un relato cronológico para facilitar su seguimiento —marcado por las efemérides más útiles para el re-

lato—, y apoyándose en una selección de ejemplos lo más familiares y próximos al lector. El resultado de este esfuerzo de generalización ha llevado a definir cuatro paradigmas en los que se basa la IA, que resultarán de gran ayuda para comprender cómo hemos llegado a donde estamos.

El objetivo de este libro es fomentar la reflexión informada y consciente sobre la evidencia de que nos encontramos ya inmersos, sin vuelta atrás, en un nuevo paradigma coevolutivo en el que humanidad y la IA se han embarcado conjuntamente, gestando una interdependencia y cohabitación que exigen respuestas sin demora, porque la IA no esperará al ser humano.

Por ello, este libro intentará responder a ocho preguntas que buscan sintetizar las principales inquietudes en torno a este tema, ordenadas de manera que compongan un relato de todas las etapas que ha vivido la IA para, finalmente, invitar al lector a reflexionar sobre qué tipo de IA queremos y adónde nos debe llevar:

1. ¿Está la IA desplazando al ser humano?
2. ¿Cómo se creó la primera IA?
3. ¿Cómo una IA empezó a aprender de los humanos?
4. ¿Cómo una IA consiguió aprender por sí misma?
5. ¿Cómo una IA ha conseguido ser creativa?
6. ¿Podrá una IA llegar a pensar?
7. ¿Nos debe preocupar el impacto de la IA en su forma actual?
8. ¿Podemos prescindir de la IA?

Creo sinceramente que todos tenemos el derecho a entender la revolución que supone la IA, pues es una de las mayores que la humanidad ha experimentado, y formarnos nuestra propia opinión. De esta manera, contaremos con las herramientas adecuadas a la hora de tomar decisiones sobre los potenciales riesgos y desafíos asociados a la IA, así como también sobre sus posibilidades y ventajas. Y podremos, además, asegurarnos de que la IA evolucione de una manera responsable y sostenible, con un impacto positivo en el futuro de la humanidad.

5.
¿Cómo una IA ha conseguido ser creativa?

El debate sobre la «creatividad» de la IA está sobre la mesa desde que se popularizaron las primeras IA capaces de generar sorprendentes imágenes a partir de un simple texto. Para muchos, la creatividad es una cualidad intrínsicamente humana que va más allá de las artes y las ciencias, como lo demostró Ferran Adrià en la gastronomía o Johan Cruyff en el fútbol. Es decir, entendemos la «creatividad» como algo más que una creación, sino también como algo innovador que transforma e inspira a otros a continuar siendo creativos.

¿Podemos hoy en día considerar creativa una IA? Está claro que su aplicación en el arte cambiará de manera contundente la naturaleza del proceso creativo. Es más, dejará de ser una herramienta de ayuda a la creación para convertirse en una compañera e, incluso, una agente creativa en sí misma.

Aunque hay quienes no dudan en considerar creativas a las IA como ChatGPT o DALL-E, otras voces, en cambio, opinan que no pueden ser tratadas de «creativas», puesto que

tan solo «generan» respuestas basadas en el aprendizaje propiciado con millones de datos creados, en última instancia, por humanos. Es por eso por lo que solo se les concede la etiqueta de «IA generativas».

IA generativa

En 2021, OpenAI lanzó una nueva IA para crear imágenes a partir de texto, bautizada como DALL-E, en un guiño al pintor Salvador Dalí. Este nuevo *software* había aprendido a partir de una gigantesca base de datos con millones de imágenes descritas en texto. Un año más tarde, la organización lanzó de manera gratuita ChatGPT, el archiconocido bot conversacional que emula la lógica del pensamiento humano en su forma comunicativa.

Los avances de las IA están conquistando ámbitos humanos, como el lenguaje o la creación artística de imágenes, mucho más rápido de lo esperado. En los últimos años hemos visto cómo las grandes compañías tecnológicas han volcado gran cantidad de recursos en el desarrollo de productos en esta dirección y están reorientando su estrategia, lanzando continuamente nuevas versiones de sus productos con capacidades cada vez superiores, que ya no incluyen solo textos de mayor calidad sino que incorporan una IA multimodal, es decir, capaces de trabajar con texto, imágenes y sonido a la vez.

La fascinación que despertó ChatGPT fue tal que en tan

solo dos meses desde su lanzamiento, a finales de 2022, sumaba ya 100 millones de usuarios únicos. Para hacerse una idea de la magnitud de lo que representa, Instagram tardó más de dos años en llegar a ese mismo número de usuarios. Estas cifras han convertido ChatGPT en la aplicación de consumo en Internet de más rápido crecimiento de la historia hasta el momento. Quizá la fascinación se debió a que, para nosotros, los humanos, el lenguaje es una ventana a la inteligencia. De cualquier forma, sea bienvenido este interés repentino por la IA de parte de la sociedad.

Cómo se llega a un bot conversacional que parece un humano

Todo empezó con la fundación en 2015 de la empresa OpenAI por parte de un grupo de líderes tecnológicos (Elon Musk entre ellos, aunque tres años después abandonó el proyecto con el pretexto de que la empresa no estaba investigando según los objetivos fundacionales, y el tiempo parece que le ha dado la razón; ahora es Microsoft la compañía que más está invirtiendo en ella). Desde entonces, OpenAI se ha centrado en construir redes neuronales cada vez de mayor tamaño, entrenadas a su vez en supercomputadores cada vez más potentes, a los que conocemos como supercomputadores a gran escala.

La primera versión de red neuronal que presentó OpenAI fue el modelo de lenguaje GPT, lanzado en 2018, que utiliza

117 millones de parámetros. Los modelos de lenguaje son un tipo de red neuronal entrenados con enormes cantidades de secuencias de letras y palabras de diferentes longitudes, y utilizan un mecanismo diferente al resto de redes neuronales: los GPT están diseñados para prestar atención a distintas partes de una frase con el fin de crear relaciones entre ellas. De esta manera, rastrea dónde aparece cada palabra o frase dentro de una secuencia, gracias a lo cual puede *interpretar* el significado de las palabras según el contexto. Estas probabilidades se calculan a partir de una selección de muchísimos textos en los que el programa busca con qué otras palabras se asocia más frecuentemente cada palabra.

En 2019, OpenaAI presentó la siguiente versión de este potente sistema de lenguaje natural, el GPT-2, compuesto ya por 1.500 millones de parámetros. Esta nueva versión estaba ya configurada de forma tal que con una breve indicación escrita (de solo una o dos frases) fuera capaz de generar una narración completa.

En mayo de 2020 llegó GPT-3, un sistema cien veces más potente que el anterior, con más 175.000 millones de parámetros. El cambio fue sustancial. Los diferentes GPT se entrenan para que sean capaces de adivinar cuál debe ser la siguiente palabra de una frase. Es decir, el modelo genera un texto palabra a palabra, ejecutándose iterativamente el algoritmo de predicción una y otra vez para cada nueva palabra. La red neuronal GPT-3 se entrenó con miles de millones de textos de diferentes fuentes de Internet, desde libros y páginas web a conversaciones reales entre usuarios. Para que el

lector se haga una idea de la dimensión: la Wikipedia entera constituye alrededor del 3 % del total de información con que se *alimentó* al nuevo programa.

El entrenamiento es el siguiente: se oculta una palabra del texto y se ejecuta la red neuronal para que la prediga. De esta manera, el esquema de entrenamiento es equivalente al del ejemplo de la red neuronal que clasificaba imágenes según aparecía un gato, o no, en ella. En este caso, sabemos cuál es la solución que buscamos porque es precisamente la palabra que se le ha ocultado al programa. Finalmente, se compara el valor que ha calculado la red neuronal con el esperado para ajustar los valores de los parámetros de la red neuronal (ver capítulo 3).

Recordemos que una interpretación simple de lo que contienen los miles de millones de parámetros de la red neuronal es su versión comprimida de todo el conocimiento que se les ha mostrado para aprender. Es un proceso similar al de comprimir un archivo. Requiere dos pasos: primero, la codificación para comprimir, durante la cual el archivo se convierte a un formato más compacto, y luego la decodificación a partir de la información comprimida, en la que se invierte el proceso. Es decir, cuando usamos estas IA para generar texto, en realidad estamos decodificando y, por tanto, la secuencia exacta de palabras que estaban en los datos originales no se encuentra almacenada tal cual, puesto que la copia comprimida solo es una representación de la información real. Sin embargo, al decodificar, es posible obtener una aproximación en forma de texto gramatical equivalente.

Esto explica algunos casos en los que las respuestas de las IA son poco acertadas, pues en cierta manera es inevitable que se haya perdido información en el proceso de *compresión*.

Es importante remarcar que estas IA generativas del estilo de GPT tienden a ser adaptables, lo que significa que pueden adquirir otras habilidades aparte de aquellas para las que fueron explícitamente capacitadas. Esto es posible gracias a su entrenamiento generalista. GPT-3, por ejemplo, no solo aprendió a escribir un texto de aspecto realista, sino que también aprendió a generar un código de programación aceptable, a pesar de que no tenía la intención explícita de hacerlo al principio.

A finales de 2022 se lanzó la versión abierta de ChatGPT (una versión mejorada de GPT-3). En esta ocasión, el programa se centró en utilizar el contenido de las conversaciones interactivas entre personas. Recordemos que el modelo del lenguaje GPT-3 solo estaba entrenado para predecir la siguiente palabra en una secuencia de texto, pero era incapaz de *comprender* su significado. En la nueva versión se mejoró el proceso de aprendizaje con la inclusión de comentarios humanos y con técnicas de aprendizaje por refuerzo, pero usando una retroalimentación con intervención humana en el ciclo de entrenamiento. ¿Por qué? Porque basar el entrenamiento en textos extraídos de Internet había tenido un efecto colateral indeseado: junto con la información válida, GPT-3 había absorbido gran parte de la desinformación y sesgos que se encuentran en la red. Por ello, para reducir la cantidad de información errónea y textos ofensivos que

producía GPT-3, hubo que ajustarla de forma «manual». Al final, el proceso de entrenamiento de las IA generativas requiere que la mano humana esté muy presente.

Estos modelos siguen evolucionando a medida que escribimos su historia. Poco después de lanzar el primer ChatGPT, OpenAI anunciaba una nueva versión, GPT4, que admite como entrada no solo texto sino también imágenes. En esta ocasión, la compañía no ha hecho público ni su entrenamiento ni detalles sobre los parámetros o requerimientos computacionales. En todo caso, representa una mejora de la versión anterior, pese a que aún es propenso a los mismos tipos de problemas de veracidad.

Poder transformacional de las IA generativas

ChatGPT todavía no es capaz de leer un libro (entendiendo por leer la facultad de comprender su contenido). La aproximación actual de cualquiera de las IA disponibles se basa en la representación de probabilidades intentando adivinar qué palabras tienden a concurrir en una frase o contexto. Es decir, se generan textos que parecen escritos por humanos, pero no significa que la IA tenga conocimiento del tema ni que haya comprendido el texto. Es algo similar al predictor de texto de WhatsApp, que nos sugiere palabras para completar el mensaje que estamos escribiendo.

Uno de los aspectos más preocupantes es la falta de veracidad y los sesgos que mencionábamos en el apartado anterior.

En muchas ocasiones la IA es capaz de responder con información falsa como si fuera cierta, bien sea porque los datos de entrenamiento no están actualizados o porque en el proceso de codificación y decodificación se ha perdido información.

No hay que perder de vista que la red neuronal no es más que un modelo limitado del mundo conformado con los valores de sus parámetros, y no un modelo del mundo real. De momento, no se ha encontrado la manera de entrenar modelos con datos extraídos de Internet sin absorber lo que se conoce como la suciedad de los datos, es decir, los bulos y contenidos ofensivos, entre otros. De ahí que, igual que hemos visto en el caso de GPT-3, la única solución que existe hasta el momento es que operadores humanos filtren la información a mano.

A pesar de todas estas limitaciones, la ola de las IA generativas no ha hecho más que empezar, y ChatGPT representa solo el primer exponente de los nuevos modelos que pronto estarán presentes en todos los aspectos de nuestras vidas. Sus posibilidades son casi infinitas: desde chatear a generar documentos sofisticados o sencillamente servir de inspiración. Como ya ha ocurrido con los traductores automáticos, las aplicaciones de creación de texto serán habituales en nuestros dispositivos y pasarán a integrarse en el *software* que utilizamos cotidianamente en nuestras rutinas productivas, tanto en el ámbito doméstico como en el empresarial y educativo.

Los gigantes tecnológicos habituales se han lanzado a una carrera para desarrollar IA generativas no solo mucho más potentes y eficientes, sino también específicas para ámbitos

determinados, es decir, entrenadas con datos personalizados (por ejemplo, medicinal o empresarial). Pronto se convertirán en las mayores expertas en el área en la que hayan sido entrenadas.

Un último apunte sobre esta carrera por la IA: la proliferación de desarrolladores ha generado también tensiones entre la comunidad de código abierto IA y las empresas privadas respecto a la exclusividad de los códigos de IA. Los primeros abogan por unas IA generativas como instrumento de creatividad e innovación abiertas, mientras que los segundos defienden su privatización.

La IA basada en fuerza bruta

¿Pero cómo hemos llegado aquí? Ya hemos dicho que el año 2012 fue un punto de inflexión para la adopción de las redes neuronales, en especial gracias al equipo de la Universidad de Toronto y su revolucionaria participación en la competición de ImageNet.

Pronto se vio que los métodos de aprendizaje de las redes neuronales podían aprovechar magníficamente técnicas de paralelismo, es decir, utilizar varios chips aceleradores de forma simultánea para reducir el tiempo de entrenamiento de las redes neuronales. Y, por último, los supercomputadores a gran escala permitieron acelerar esto aún más, interconectando una gran cantidad de máquinas con varios chips aceleradores cada una.

Sin duda, el paralelismo es una técnica capital en la supercomputación a gran escala. Pongamos como ejemplo la red neuronal de Google para traducción multilingüe. Consiste en una red con 600 mil millones de parámetros, es decir, una capacidad de computación equivalente a 22 años si solo se dispusiera de un chip acelerador tipo TPU (como el que se usó en AlphaZero). Pero dado que el sistema de Google utiliza 2.048 chips de este tipo en simultáneo, consigue realizar el entrenamiento en solo cuatro días.

En el último año, las necesidades de computación para entrenar las IA generadoras de texto se han multiplicado por dos cada tres o cuatro meses, con lo que las infraestructuras con gran capacidad de computación se han revelado como fundamentales. Hoy en día es inconcebible pensar en un supercomputador a gran escala que no cuente con un *hardware* pensado para entrenar una IA. Uno de los más recientes es el MareNostrum 5 (ha entrado en funcionamiento durante 2023), que incluye 4.480 chips aceleradores GPU de última generación fabricados por Nvidia. Es uno de los nodos principales de la red europea de supercomputación EuroHPC.

La gran capacidad computacional disponible en la actualidad ha permitido a la comunidad de IA avanzar los últimos años con mucha rapidez y diseñar redes neuronales cada vez más complejas, aunque esto ha exigido aumentar la infraestructura de computación a niveles nunca vistos. Estamos inmersos en lo que podríamos llamar el paradigma de la IA basada en «fuerza bruta». Es decir, algoritmos de miles de millones de parámetros que necesitan supercomputadores a

gran escala para ser entrenados con ingentes cantidades de datos. Unos recursos, desde luego, solo al alcance de muy pocos.

En resumen,
¿cómo una IA ha conseguido ser creativa?

- ChatGPT fue una IA generativa orientada a la comprensión y generación de diálogos para conversar con los humanos basada en un tipo de red neuronal de miles de millones de parámetros.
- Este tipo de redes puede adquirir habilidades mucho más complejas de aquellas para las que fueron capacitadas debido a su entrenamiento generalista con grandes conjuntos de datos.
- Para evitar absorber en lo posible la desinformación y los sesgos que contiene Internet, se ha mejorado el entrenamiento de las redes neuronales con técnicas de aprendizaje por refuerzo en el que intervienen equipos humanos.
- La ola de IA generativas no ha hecho más que empezar. ChatGPT solo representa el primer exponente de unas IA generativas que estarán pronto presentes en todos los aspectos de nuestras vidas y conllevará un impacto social sin precedentes.
- La IA generativa está basada en el paradigma de fuerza bruta, un escenario en el que redes neuronales de miles de millones de parámetros deben ser entrenadas con grandes cantidades de datos, para lo que se requieren supercomputadores a gran escala.

Su opinión es importante.
En futuras ediciones estaremos encantados
de recoger sus comentarios sobre este libro.

Por favor, háganoslos llegar a través de nuestra web:

www.plataformaeditorial.com

Para adquirir nuestros títulos,
consulte con su librero habitual.

«*I cannot live without books*».
«No puedo vivir sin libros».
THOMAS JEFFERSON

Desde 2013, Plataforma Editorial planta un árbol
por cada título publicado.