¿Qué es la inteligencia artificial?

RAMÓN LÓPEZ DE MÁNTARAS Y PERE BRUNET

La inteligencia artificial (IA en adelante) es una colección de componentes computacionales que permiten construir sistemas que emulan funciones realizadas por el cerebro humano.[1] El campo de la IA comenzó a mediados de los años cincuenta y desde entonces ha pasado por ciclos de promesas, entusiasmo, críticas y dudas.

Podemos distinguir entre IA basada en conocimiento e IA basada en datos. La IA basada en el conocimiento, que empezó a desarrollarse a finales de los años setenta, intenta modelar el conocimiento humano mediante modelos informáticos. Comienza de arriba hacia abajo a partir de un análisis humano sobre qué conceptos y conocimientos utilizan los individuos para resolver problemas o responder consultas en un dominio concreto de especialización, formalizando e implementando dichos conocimientos mediante lenguajes de representación basados en la lógica matemática. Esta IA basada en el conocimiento utiliza bases de conocimientos, modelos conceptuales, ontologías, estrategias de razonamiento automatizado, técnicas heurísticas de resolución de problemas y aprendizaje profundo.[2]

En cambio, la IA basada en datos se ha desarrollado mayoritariamente a partir del siglo XXI. Comienza de abajo a arriba a partir del análisis de grandes cantidades de datos que se procesan mediante algoritmos estadísticos de aprendizaje, tales como los algoritmos de aprendizaje profundo, para extraer patrones en dichos datos que se usan para resolver

[1] Luc Steels y Ramón López de Mántaras, «The Barcelona declaration for the proper development and usage of artificial intelligence in Europe», *AI Communications* 31, 2018, pp. 485-494: https://content.iospress.com/articles/ai-communications/aic180607 . Véase también el texto original de la declaración de Barcelona en: https://www.iiia.csic.es/barcelonadeclaration/

[2] Ramón López de Mántaras (2018), *op. cit.*

A fondo

problemas cuya solución se construye en base a los patrones extraídos.[3] La IA basada en datos requiere una cantidad ingente de datos de entrenamiento, así como computación de altas prestaciones para poder funcionar. Además, los datos de entrenamiento tienen que ser de alta calidad para que dicho funcionamiento sea correcto.

La IA basada en conocimiento ha demostrado ser muy eficiente en tareas que requieren razonamiento o planificación, mientras que la IA basada en datos funciona mejor en tareas que, en lugar de razonamiento o planificación, requieren sobre todo detectar patrones estadísticos como por ejemplo el procesamiento de imágenes o lenguaje. Pero, con toda probabilidad vamos a ver cada vez más aproximaciones híbridas que combinen ambos enfoques.[4]

Uno de los desarrollos sorprendentes de la IA basada en datos nos llegó de la mano de los traductores automáticos. El investigador Franz Josef Och fue pionero en diseñar los primeros algoritmos en 2003, y luego Google los incorporó entre los años 2005 y 2007. Estos nuevos traductores funcionaban tras aprender de ingentes cantidades de datos. Según Och,[5] para poder traducir bien entre dos idiomas se necesita un corpus de texto bilingüe de más de 150 millones de palabras y dos *corpus* monolingües de más de mil millones de palabras.

En 2012, un equipo de la Universidad de Toronto liderado por Geoffrey Hinton consiguió que un tipo de red neuronal, llamada «convolucional», alcanzara un 85% de aciertos al clasificar, entre mil categorías posibles, 150.000 imágenes de la base de datos ImageNet. Tanto estas redes de clasificación como las de los traductores automáticos son casos concretos de redes neuronales de aprendizaje profundo, ejemplos de la IA basada en datos. La idea proviene de los trabajos del investigador japonés Kunishiko Fukushima en 1980, quien había desarrollado el «neocognitrón», una red neuronal artificial inspirada, a su vez, en los estudios de David Hubel y Torsten Wiesel sobre el sistema visual de los animales, trabajos por

[3] Ramón López de Mántaras (2018), *op. cit.*

[4] Ramon López de Mántaras, *100 coses que cal saber sobre intel·ligència artificial*, (en catalán), Cossetània, Barcelona, 2023, pp.. 48 a 50.

[5] Josef Franz Och, «Statistical Machine Translation: From Single-Word Models to Alignment Templates», Technical Report, RWTH Aachen, Department of Computer Science, 2003, disponible en: http://www-i6.informatik.rwth-aachen.de/publications/download/520/OchF.J.—StatisticalMachineTranslationFromSingle-WordModelstoAlignmentTemplates—2002.pdf ; también su presentación de 2005, ya como empleado de Google: «Machine Translation», Summit 2005, Phuket, 2005, disponible en: http://www.mt-archive.info/MTS-2005-Och.pdf

los que en 1981 estos investigadores recibieron el premio Nobel.[6] Hubel y Wiesel descubrieron que nuestra corteza visual se encuentra organizada según una jerarquía de capas, de tal manera que las neuronas contenidas en cada capa detectan características de complejidad creciente en los objetos de una imagen.

En estos y otros casos, estas redes deben entrenarse primero con una enorme cantidad de datos. De hecho, la IA basada en datos trabaja en dos fases: la primera de aprendizaje o entrenamiento y la segunda, de uso (en algunos casos, ambas fases interaccionan de manera que los sistemas continúan aprendiendo durante su uso). La primera es altamente costosa y laboriosa y requiere gran potencia de cálculo, mientras que la segunda puede ejecutarse en ordenadores personales o teléfonos móviles y es eficiente y rápida.

Hay que observar que, por ejemplo, en el caso de la clasificación de imágenes, hasta hace poco ni había bases de datos de imágenes lo suficientemente grandes ni existía la potencia de cómputo necesaria para poder entrenar redes multicapa en un tiempo razonable. Dicho entrenamiento consiste en ajustar los valores numéricos correspondientes a los “pesos” de las conexiones que unen las neuronas artificiales de la red. Para ello, a la máquina se le proporciona una gran cantidad de imágenes ya etiquetadas, y un algoritmo va ajustando los valores de los pesos en función de los errores que comete la red al clasificar las imágenes de entrenamiento. Antes de comenzar el entrenamiento los valores asignados a las conexiones son aleatorios, y el proceso finaliza cuando los pesos alcanzan valores estables. Por supuesto, todo ello requiere partir de una representación numérica de la imagen (o del texto en el caso de la traducción automática). Esto se consigue asociando un número a cada píxel (o un conjunto de números a cada frase), de modo que, desde el punto de vista de la máquina, las imágenes y los textos no son más que un enorme conjunto de números.[7]

Para la segunda fase, una vez la red neuronal ya ha sido entrenada, disponemos de una inmensa estructura de neuronas artificiales (pequeños elementos de *software* dispuestos en capas, cada uno de los cuales calcula su ”valor“ promediando muchos de los valores de las neuronas de la capa anterior). La red neuronal, con-

[6] Ramón López de Mántaras, «El traje nuevo de la inteligencia artificial», Investigación y ciencia, Julio de 2020, disponible en: https://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/una-nueva-era-para-el-alzhimer-803/el-traje-nuevo-de-la-inteligencia-artificial-18746

[7] Para más detalle, véase Ramón López de Mántaras 2020, *op. cit.*

junto de neuronas con su estructura conectiva y conjunto de ”pesos“ asociados a las conexiones que las unen, puede ya almacenarse como cualquier otro fichero y exportarse a los ordenadores o teléfonos móviles que la usarán. Luego, en esta segunda fase de uso, los datos concretos (un texto, una imagen o aquello que requiera la tarea que el usuario desea resolver) se convierten a una representación numérica que alimenta la primera capa de neuronas. La información se va propagando capa a capa a través de los pesos asociados a las conexiones y finalmente, los valores asociados a las neuronas de la última capa acaban conformando la respuesta del sistema de IA.[8]

Estos sistemas de IA basada en datos (sistemas de IA en lo que sigue) han experimentado un auge espectacular en los últimos años, con aplicaciones que van desde el diagnóstico precoz en medicina y la predicción del plegado de las proteínas hasta la robótica moderna, pasando por campos tan diversos como los juegos por ordenador, la previsión del impacto del cambio climático, el juego del tenis, la investigación en coches autónomos o las armas autónomas.[9]

Con todo, y a pesar de los éxitos del aprendizaje profundo aplicado al procesamiento del lenguaje, vemos que, contrariamente a lo que ha llegado a afirmarse, seguimos estando muy lejos del nivel humano. La razón de dichas exageraciones seguramente obedece a la competencia entre empresas para hacerse con la parte más grande de un pastel que es extremadamente lucrativo. Pero, aunque aún falte mucho para lograr traducciones automáticas de calidad similar a las de un humano profesional, no cabe duda de que una herramienta como Google Translate resulta muy útil si no somos muy exigentes con el resultado y si supervisamos y corregimos el resultado final.

El esfuerzo por llegar a construir máquinas que de algún modo se asemejen a nosotros ha generado sistemas que se equivocan como nosotros

En todo caso, a menudo ni siquiera los diseñadores de los sistemas de aprendizaje profundo saben con exactitud por qué la máquina funciona cuando acierta ni por

[8] Esta es una explicación simplificada. Los sistemas de IA pueden contener varias redes neuronales y pueden contemplar aprendizaje dinámico de manera que el sistema continúe aprendiendo a partir del uso de la red, en caso de que haya forma de verificar si las respuestas que va dando son o no correctas (si no lo son, el sistema modifica los pesos de las conexiones entre neuronas para intentar evitar que el error detectado se repita en el futuro). En este caso, las dos fases de aprendizaje y uso no son independientes.

[9] Para más detalle, véase Ramon López de Mántaras, 2023, *op. cit.*, pp. 56 a 265.

qué falla cuando se equivoca. Este serio inconveniente, conocido como «problema de la caja negra», hace que sea prácticamente imposible explicar las decisiones que toman estos sistemas.[10] Y es que los sistemas de IA cometen errores. El esfuerzo por llegar a construir máquinas que de algún modo se asemejen a nosotros ha generado sistemas que se equivocan como nosotros. Lo vemos en los sistemas de traducción automática y en muchos otros. El porcentaje de error depende del tipo de problema, de la calidad de los datos de aprendizaje, de la estructura de la red neuronal y de la calidad del proceso de entrenamiento, pero nunca es nulo. Es algo que no tiene porqué ser grave en muchos casos, si repasamos el resultado del sistema y lo corregimos en caso necesario (como hacemos cuando usamos los sistemas de traducción) o en los sistemas que generan hipótesis que luego vamos refrendando, o en los sistemas de AI que a pesar de sus errores funcionan en promedio (sistemas publicitarios en los que lo que cuenta es los clientes que captemos, a pesar de que en otras personas no funcionen). Pero esto conlleva que en aplicaciones críticas como pueden ser las de diagnóstico médico o las militares y de control y vigilancia, la postsupervisión por parte de una persona experta que se haga responsable de la decisión final sea imprescindible.

Por otra parte, la IA en realidad no es inteligencia en el sentido comúnmente aceptado. Lo que poseen los sistemas de IA son habilidades para resolver problemas y tareas específicos, pero sin ningún tipo de comprensión sobre la naturaleza de los elementos con los que trabaja y sobre sus interrelaciones. Su falta de sentido común les hace capaces de identificar una persona que está de pie delante de una pared sin saber que es una persona y que esta no puede atravesar la pared.[11]

En este momento vemos una fuerte ola de adopción entusiasta de la IA en muchas áreas de la actividad humana. Pero la ausencia de conocimientos de sentido común imposibilita que los sistemas de IA puedan comprender ni el lenguaje ni lo que “perciben” sus sensores. Del mismo modo, no pueden gestionar situaciones imprevistas ni tampoco aprender a partir de la experiencia. Los sistemas de IA basados en aprendizaje profundo pueden aprender correlaciones entre eventos (funciones matemáticas simétricas) pero no las relaciones asimétricas que nos llevan a diferenciar causas de efectos. Pueden asimilar, por ejemplo, que la salida del sol está relacionada con el canto del gallo, pero no que la primera es causa del segundo, y no al revés. El aprendizaje de las relaciones causa-efecto por parte

[10] Véase Ramón López de Mántaras 2020, *op. cit.*
[11] Ramon López de Mántaras, Cossetània 2023, *op. cit.*, página 45.

de los sistemas de IA es justamente una línea de investigación actual muy interesante.[12]

Como ya dijo Arthur Clarke en los años sesenta, cualquier tecnología que sea suficientemente sofisticada no puede distinguirse de la magia.[13] De aquí es de donde surge una buena parte del relato social que actualmente rodea los sistemas de IA. No entendemos porqué un sistema de IA puede traducir textos o responder acertadamente a preguntas que le hacemos, de la misma manera que nuestros abuelos no podrían entender cómo podemos, con un simple teléfono móvil, mandar fotos al instante a cualquier punto del planeta. La sorpresa ante el hecho de no saber entender estos sistemas nos lleva a considerarlos mágicos. Y la magia nos transporta al campo de la ficción y los mitos. Abandonamos la realidad y, ya instalados en el ámbito de los mitos, creemos que el potencial de la IA no tiene límites y que estos sistemas nos llevarán a inteligencias superiores a la humana.[14]

La ausencia de conocimientos de sentido común imposibilita que los sistemas de IA puedan comprender ni el lenguaje ni lo que "perciben" sus sensores

La fascinación se amplifica porque llueve sobre nuestra innata tendencia a generar mitos y a disfrutar de ellos. Creamos máquinas y soñamos pensando que nos dominarán. Pero nuestro deber es separar los mitos de la realidad. Podemos inventar grandes historias sobre lo que nos puede deparar la IA, pero debemos dejarlas en el rincón de los mitos y, en cambio, escuchar a los expertos para saber cuál va a ser la realidad. Michael Shermer habla de la imposibilidad de que lleguemos a ver máquinas que piensen, que sean autoconscientes y que tengan emociones. Este apocalipsis, esta singularidad, dice irónicamente, lo más probable es que nos llegue en algún momento entre los años 2525 y 9595.[15]

Pero por muy sofisticada que llegue a ser la IA en el futuro, siempre será diferente de la humana. Porque el desarrollo mental humano se nutre de las interacciones

[12] Ramon López de Mántaras, «Intel·ligència artificial versus intel·ligència humana» (en catalán), en *IA: Inteligencia Artificial*, catálogo de exposición, CCCB, 2023, pàgina 48. Disponible en: https://www.cccb.org/es/publicaciones/ficha/ia-inteligencia-artificial/243181

[13] Ramon López de Mántaras 2023, *op. cit*., página 47.

[14] Ramon López de Mántaras, Cossetània 2023, *op. cit*., página 44.

[15] Pere Brunet, Tica Font y Joaquín Rodriguez, *Robots Asesinos: 18 preguntas y respuestas*, Centro Delàs de Estudios para la Paz, 2021, p. 01.2, disponible en: https://centredelas.org/wp-content/uploads/2021/12/RobotsAsesinos_18PreguntasYRespuestas_DEF.pdf . Ver también: https://centredelas.org/robots-asesinos-18-preguntas-y-respuestas/?lang=es

Imagen generada por la inteligencia artificial Mid Journey en que, a modo de experimento, se le pidió que se retratara a sí misma en base a estas preguntas: Qué es la IA, riesgos y oportunidades, impactos socioecológicos.

con el entorno, que a su vez se basan tanto en la corporeidad de nuestros sentidos y sistema perceptivo como en nuestro sistema motor. Una corporeidad perceptiva y motora que no existe en las inteligencias artificiales no corpóreas. Junto con la intencionalidad, esencialmente humana, que no tienen ni tendrán los sistemas de IA,[16] que por ello ven necesariamente limitadas sus capacidades de aprendizaje.[17] En todo caso, la fascinación no cesará. Y será perfectamente aceptable si sabemos mantenerla en el ámbito de los mitos mientras, al mismo tiempo, nos esforzamos por entender la realidad y los hechos objetivos. Porque quienes nos querrán controlar serán personas concretas, no máquinas.[18] Y porque los verdaderos problemas de la IA no provienen de una supuesta singularidad tecnológica que pueda surgir de hipotéticas y futuras superinteligencias artificiales. Los verdaderos problemas están en la manipulación, en el uso ilícito de datos privados y en su privacidad, en la vigilancia y el control masivo de la ciudadanía, en la autonomía de sistemas que pueden usarse contra las personas (como las armas autónomas), en la confianza excesiva en las capacidades de la IA, en los sesgos de los algoritmos, en la imposibilidad de rendición de cuentas en el caso de funcionamiento erróneo, y en el excesivo poder que acumulan unas pocas empresas tecnológicas.[19]

Los verdaderos problemas están en la posibilidad de manipulación, en el uso ilícito de datos privados y el control masivo de la ciudadanía

En 2020, las investigadoras Timnit Gebru y Margaret Mitchell, codirectoras en aquel momento del equipo de ética de Google, ya advirtieron del riesgo que suponía que la gente asignase intención comunicativa y comprensión del lenguaje a artefactos.[20] Tras haber publicado esta consideración ética, Google las despidió.

Dado el interés público en la IA y el entusiasmo de muchas organizaciones, tanto privadas como instituciones gubernamentales, por desarrollar aplicaciones que afecten a las personas en su vida diaria, es importante que la comunidad de IA, incluidos los desarrolladores de aplicaciones así como los investigadores, participen en debates abiertos, en parte para evitar expectativas excesivas con una reacción inevitable posterior y en parte para evitar un uso inadecuado de la IA que

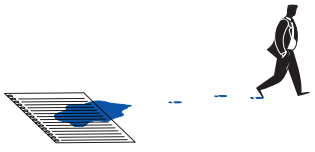
[16] Ramon López de Mántaras 2023, *op. cit*., p. 52.
[17] Ramon López de Mántaras 2023, *op. cit*., p. 51.
[18] Pere Brunet, Tica Font y Joaquín Rodriguez, *op. cit*., p. 01.2
[19] Ramon López de Mántaras 2023, *op. cit*., p. 52.
[20] Ramon López de Mántaras 2023, *op. cit*., p. 49.

puede causar efectos secundarios negativos innecesarios y sufrimiento humano indebido. Al mismo tiempo, debemos darnos cuenta de que ningún conjunto de reglas o limitaciones tecnológicas incorporadas puede evitar el uso malicioso por parte de actores sin escrúpulos.[21] La responsabilidad final siempre recae en los seres humanos, tanto como diseñadores como usuarios, y deben rendir cuentas.

Ante los evidentes peligros a que nos enfrenta un desarrollo de la IA todavía no regulado y basado en el lucro de unas pocas corporaciones, muchos expertos están pidiendo una regulación que garantice que estos sistemas vayan dirigidos a cubrir necesidades de las personas, respetando sus derechos y sin dañarlas, violentarlas, controlarlas o manipularlas. Pero, además, es imprescindible educar a los ciudadanos (en particular a los jóvenes en las escuelas y universidades y a los políticos) sobre los beneficios y riesgos de estas tecnologías de IA. Los estudiantes de ciencias e ingeniería deben recibir una formación ética que les permita entender las implicaciones sociales de las tecnologías que desarrollarán.[22] Y los ciudadanos en general deben exigir estar mejor informados, desde un sentido crítico que les permita discernir, que les aporte mayor capacidad para evaluar los riesgos tecnológicos y que lleve a hacer valer sus derechos. Las administraciones deben ser valientes para regular y visionarias para invertir en una educación que capacite adecuadamente a sus jóvenes y ciudadanos.

Ramón López de Mántaras Badia es profesor de investigación del CSIC y director del Instituto de Investigación en Inteligencia Artificial. Es uno de los pioneros de la inteligencia artificial en España.

Pere Brunet i Crosa es doctor y catedrático jubilado de la Universidad Politécnica de Catalunya, investigador del Centre Delàs d'Estudis per la Pau y divulgador científico.



[21] Luc Steels y Ramón López de Mántaras, *op. cit.*
[22] Ramon López de Mántaras 2023, *op. cit*., página 52.