Reseña de *The Road to Conscious Machines. The Story of AI*

Autor:
Michael Wooldridge. Profesor de Ciencias de la Computación y Director del Departamento de Ciencias de la Computación de la Universidad de Oxford

Editorial:
Colección Pelican Books. Penguin Random House

Eva Valentí Ramírez
Jefe de Departamento de Revisión Actuarial, Dirección de Gestión de Riesgos
Consorcio de Compensación de Seguros

Es frecuente encontrar en los medios de comunicación, en el cine o en redes sociales ideas del tipo:

1. La inteligencia artificial (IA) nos dejará a todos en paro en un futuro próximo. Realizará cualquier tarea mejor que nosotros sin necesidad de recibir un salario.

2. A los gobiernos o a determinados grupos de poder les resultará fácil manipular a los ciudadanos mediante noticias falsas construidas con imágenes y sonidos obtenidos por IA.

3. Pronto se fabricarán máquinas superinteligentes, capaces de mejorarse a sí mismas a toda velocidad, de manera que evolucionarán por su cuenta y quedarán fuera de control.

Tales afirmaciones se repiten tan a menudo que mucha gente ha acabado aceptándolas como ciertas aunque, desde el punto de vista técnico, no tengan fundamento.

En otros casos, en vez de predecir grandes catástrofes, se exhibe una confianza exagerada en las capacidades de la IA como, por ejemplo, en esta otra, aplicable al sector asegurador:

La IA busca construir máquinas que «imiten» cada vez más fielmente el comportamiento humano, con el objetivo final de hacerlas indistinguibles de nosotros. Máquinas con el mismo rango de capacidades que tiene la inteligencia humana, lo que se conoce como **IA general**, es decir, una inteligencia autónoma y auto-consciente, con capacidad para hacer planes, para razonar, para mantener una conversación, entender los chistes y narrar historias... Y esto no se ha conseguido aún.

«La implantación de herramientas de IA supondrá una revolución para la industria aseguradora. Gracias a ellas se mejorará la experiencia del cliente y la eficacia en la comercialización y gestión de siniestros de una manera rapidísima. Las aseguradoras experimentarán un crecimiento sin precedentes».

Para Michael Wooldridge –autor del libro–, que se define a sí mismo como *perteneciente a la primera generación de humanos que jugó con un ordenador en su adolescencia*, la IA no es tan poderosa ni está tan avanzada como pensamos.

Ha escrito este libro para tratar de poner las cosas en su sitio y dar una versión más realista de lo que es –y no es– la IA, cómo ha acumulado logros técnicos desde sus inicios en los años 50 y en qué punto se encuentra actualmente.

De manera rigurosa, pero entretenida y con múltiples ejemplos para facilitar la comprensión, el autor va presentando los conceptos básicos de la IA y explica cómo se ha ido convirtiendo en la potente disciplina que hoy conocemos.

Sabemos que los ordenadores son brillantes haciendo tareas concretas, las realizan sin errores y a toda velocidad. Un ordenador de sobremesa puede hacer en 1 segundo lo que a una persona, trabajando sin descansar y sin equivocarse, le supondría 3.700 días de trabajo.

En la próxima década veremos coches totalmente autónomos, traductores simultáneos de calidad, programas capaces de detectar diferencias mínimas en los píxeles de una radiografía, localizando tumores con una eficacia muy superior a la de un médico. Incluso habrá aplicaciones para el móvil que podrán detectar síntomas de demencia por la manera en que su propietario lo utiliza.

Todo lo anterior nos facilitará enormemente la vida. Pero ese no es el objetivo último de la IA.

La IA busca construir máquinas que «imiten» cada vez más fielmente el comportamiento humano, con el objetivo final de hacerlas indistinguibles de nosotros. Máquinas con el mismo rango de capacidades que tiene la inteligencia humana, lo que se conoce como **IA general**, es decir, una inteligencia autónoma y auto-consciente, con capacidad para hacer planes, para razonar, para mantener una conversación, entender los chistes y narrar historias… Y esto no se ha conseguido aún.

Los científicos han avanzado por callejones sin salida antes de caer en la cuenta de que tenían que retroceder y volver a empezar por otro camino. Aún desconocen, incluso, si la IA general es viable y no hay consenso, tampoco, en si es deseable.

La primera parte del libro cuenta su historia.

Todo empieza, no podría ser de otra manera, con Alan Turing y el «problema de decisión», *Entscheidungsproblem* en alemán, que fue el primer paso dado en el desarrollo de la IA cuando esa disciplina ni siquiera tenía nombre ni existía comunidad científica alguna dedicada a ella.

Los problemas de decisión son problemas matemáticos que se resuelven con un sí o un no, por ejemplo: ¿es 2+2=4? Un problema de decisión es decidible si puede resolverse siguiendo unos pasos fijos (una receta); es decir, un ordenador podría resolverlo en un tiempo finito. La cuestión es: ¿son todos los problemas de decisión decidibles o algunos de ellos no pueden resolverse siguiendo unos pasos fijos?; es decir, un ordenador, por muy veloz que fuera ejecutando instrucciones, tardaría un tiempo infinito. Para responder a esta pregunta Alan Turing construyó la «máquina de Turing».

En esa primera etapa, entre 1956 y 1974, conocida como la Era Dorada, todo parecía posible. Fue una época de optimismo desmedido. Se daban nombres extravagantes a los sistemas desarrollados; los científicos tenían que trabajar de noche porque los ordenadores se usaban para actividades más productivas en las horas normales de oficina. Se trataba de construir robots que pudieran mantener algo parecido a una conversación o realizar tareas prácticas como ordenar un almacén. Pero a mediados de los 70, se vio que, tras dos décadas de investigación, solo se habían hecho avances muy básicos y una parte de la comunidad científica empezaba a considerar la IA como una pseudociencia.

La IA cayó entonces en un oscuro periodo de estancamiento, hasta que la investigación cambió de dirección y se empezaron a desarrollar los primeros **sistemas expertos**. Construir un sistema experto consiste en dotar al ordenador de conocimientos para realizar tareas específicas, conocimientos que las personas expertas en dichas

tareas solo adquieren tras un largo periodo de entrenamiento, y que la máquina desempeña de manera mucho más eficaz que un humano. Por primera vez se vislumbró que la IA podía aportar beneficios económicos.

Así que, a finales de los años 70 comenzó de nuevo la euforia. Pero al terminar la década de los 80 no se habían conseguido avances reseñables; resultó que no era tan fácil traducir la experiencia humana a instrucciones codificadas para que las ejecutara un ordenador. La comunidad científica dedicada a la IA fue de nuevo acusada de vender humo, prometer mucho y no llegar a nada concreto.

Una vez más se produjo un cambio de orientación de la investigación en IA que tendría ocupados a los científicos de la siguiente década –1985-1995–. Se llegó a la conclusión de que sólo se podría avanzar si los sistemas adquirían directamente la información del entorno real donde se encontraran. Se trataba de establecer los comportamientos que el sistema debía exhibir en cada situación, organizándolos en capas jerárquicas, de manera que se primara uno u otro –**IA conductual**–. El siguiente paso fue el desarrollo de **agentes**: sistemas de IA completos; es decir, autónomos y capaces de ejecutar las tareas encomendadas por sus usuarios de manera integral.

Mientras tanto, y desde los inicios de la IA, la investigación avanzaba también por otro revolucionario camino: fabricar máquinas capaces de aprender.

El objetivo de construir computadores con capacidad de aprendizaje es diseñar programas que obtengan unos resultados a partir de unos datos de entrada, sin que el programa incluya explícitamente la «receta» para llegar a ellos.

Para eso es necesario «entrenar» al programa. Existen dos tipos de aprendizaje. El primero, el **aprendizaje supervisado**, se consigue proporcionando a la máquina la colección más variada de situaciones posibles. Aquí surge uno de los problemas éticos a los que se enfrenta la IA: si el conjunto de datos aportados para el entrenamiento está sesgado, las decisiones que tomará el ordenador reproducirán ese sesgo, creando situaciones injustas.

Por ejemplo, un banco que utilice un programa para detectar el riesgo de cada cliente en la concesión de créditos bancarios. Normalmente, un programa de ese tipo se entrena a partir de una colección de registros de clientes antiguos, etiquetados con una clasificación de riesgo alto o bajo. Pero, dado que trabajar con demasiados datos de cada cliente hace el aprendizaje muy lento, ¿qué datos hay que omitir si se desconoce cuáles son relevantes para determinar el riesgo? Por ejemplo, si el único dato que se aporta para el aprendizaje es la dirección del cliente, es muy posible que esto lleve al programa a discriminar a los residentes en determinados barrios, impidiendo a potenciales buenos clientes tener acceso a un crédito.

En el segundo tipo de aprendizaje, **aprendizaje de refuerzo**, no se le dan al programa datos explícitos, se le deja tomar decisiones al azar que reciben un *feedback* negativo o positivo según resulten ser malas o buenas. El programa tiene en cuenta ese *feedback* al tomar la siguiente decisión.

Uno de los desafíos actuales es evitar los sesgos en los algoritmos, porque el algoritmo decide sin que sepamos cuál es el camino por el que llega a las conclusiones a las que llega.

Así pues, para que la máquina tome decisiones solo hay que decirle cómo debe hacerlo y la propia máquina cambiará su comportamiento, es decir: aprenderá.

Pero, ¿cómo aprende un programa? La técnica de aprendizaje –*deep learning*– consiste en dotar al ordenador de una arquitectura de **redes neuronales** que pueden ser entrenadas. Esta estructura está inspirada en el sistema nervioso de los animales, donde el impulso nervioso es transmitido por cada neurona a la siguiente, que se activa o no en función de los neurotransmisores segregados en la sinapsis. En el ordenador, esos neurotransmisores

químicos son sustituidos por dos números: el **peso y el umbral de activación**; según las combinaciones entre sus valores se activa o no la siguiente neurona de la red.

La parte final del libro está dedicada a revisar el presente y el futuro de la IA.

Se analizan dos logros de la IA que ya son una realidad: los coches sin conductor y la aplicación de la IA en el control de la salud.

En dos capítulos, titulados con ironía: *Lo que imaginamos que podría ir mal y Cosas que realmente podrían ir mal*, para remedar nuestro miedo a «lo nuevo» y nuestras injustificadas certezas, se desgranan riesgos reales de la IA a los que conviene prestar atención. Se reflexiona sobre el futuro del trabajo y de los derechos humanos más allá del tópico de trabajadores alienados dirigidos por un algoritmo. Se hacen interesantes especulaciones sobre los cambios que la IA provocará en la sociedad basada en el trabajo tal y como lo conocemos hoy. Se reflexiona también sobre el problema de las noticias falsas y el dilema moral planteado por las armas autónomas, dirigidas por una IA que toma sus propias decisiones.

Como punto final, el autor se divierte fantaseando sobre cómo sería una máquina que realmente fuera indistinguible de un ser humano. ¿En qué consiste la consciencia?, ¿podremos saber si tales máquinas son realmente autoconscientes?

¿Y si la IA general, simplemente, no es posible?