Introducción

"La mayor parte de lo que hacemos con el aprendizaje automático ocurre bajo la superficie. Aunque no sea visible, gran parte del impacto del aprendizaje automático será así: una mejora silenciosa pero significativa de las operaciones esenciales".

Jeff Bezos[1]

"La inteligencia artificial (AI) es el campo de la ciencia y la ingeniería centrado en crear máquinas inteligentes, y en especial programas informáticos inteligentes. Está relacionada con la tarea similar de utilizar ordenadores para comprender la inteligencia humana, pero la AI no tiene por qué limitarse a métodos biológicamente observables"[2].

Esta fue la definición de AI que ofreció John McCarthy, profesor de la Universidad de Stanford, uno de los fundadores de esta disciplina y coautor del término "inteligencia artificial".

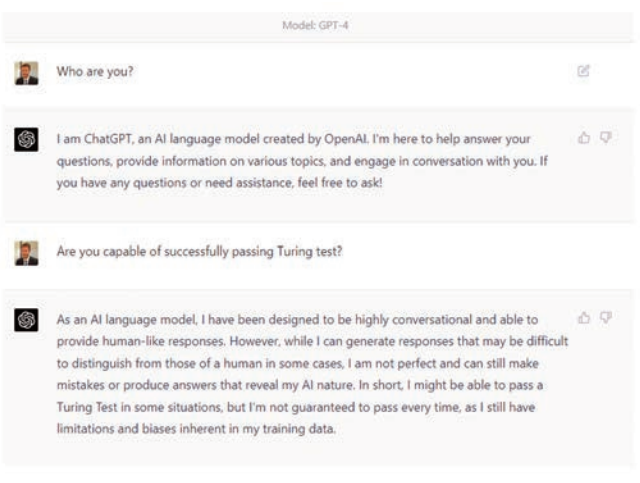
Sin embargo, ya en 1950 Alan Turing se preguntaba[3]: "¿pueden las máquinas pensar?", y formulaba lo que más tarde se conocería como "test de Turing": una prueba de la capacidad de una máquina para mostrar una inteligencia indistinguible de la de un ser humano. Turing propuso que un evaluador humano juzgara las conversaciones en lenguaje natural entre una persona y una máquina diseñada para generar respuestas similares a las humanas. Si el evaluador no era capaz de distinguir la máquina del humano, la máquina habría superado la prueba.

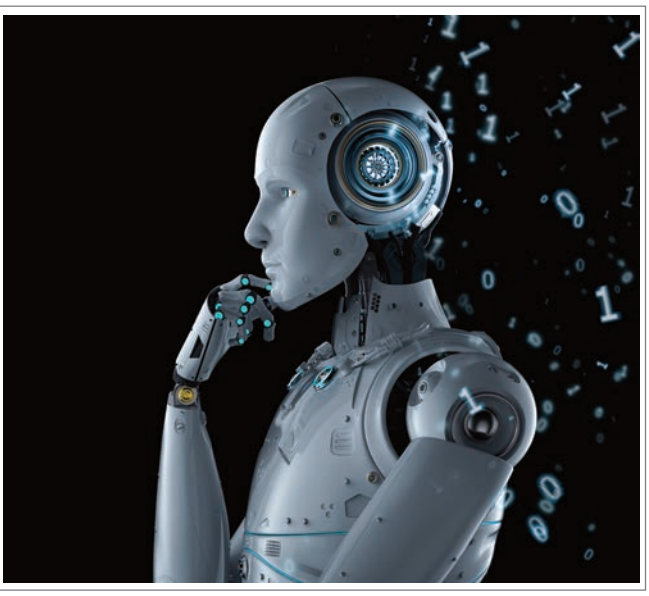
Aunque hay controversia al respecto[4], muchos autores consideran que ya hay inteligencias artificiales que podrían superar el test de Turing, como GPT-4, de la Open AI Foundation, aunque la misma GPT-4 no lo tiene totalmente claro (Fig. 1). Asimismo, existen tests más sofisticados, como la prueba de esquemas de Winograd, que consiste en la resolución de anáforas complejas que requieren conocimiento y sentido común[5], algo que la AI actual no parece ser capaz de hacer aún.

[1]Bezos (n. 1964), J., fundador, presidente ejecutivo y ex CEO de Amazon.
[2]McCarthy (2004). Profesor de Informática en la Universidad de Stanford.
[3]Turing (1950). Matemático, lógico, informático teórico, criptógrafo, filósofo y biólogo teórico británico.
[4]Harnad (2003). Catedrático de Psicología en la Universidad de Quebec en Montreal (UQAM) y en la Universidad McGill, y profesor emérito de Ciencias Cognitivas en la Universidad de Southampton.
[5]Un esquema de Winograd es una pregunta de opción binaria donde (i) hay dos partes mencionadas en la pregunta; (ii) se utilizan pronombres para referirse a ellas; (iii) hay una ambigüedad sobre a quién se refiere el pronombre; y (iv) hay palabras específicas que pueden cambiar la respuesta correcta. En un ejemplo del mismo Terry Winograd (profesor de Ciencias de la Computación en la Universidad de Stanford):
- Pregunta: los concejales de la ciudad negaron a los manifestantes la autorización porque [temían/defendían] la violencia. ¿Quién [teme/defiende] la violencia?
- Respuesta: [los concejales / los manifestantes].

Con ello, se puede generar un test alternativo al Test de Turing, utilizando dichas preguntas y penalizando fuertemente las respuestas erróneas (véase Levesque (2014)).

Figura 1. Conversación con GPT-4 sobre su capacidad para pasar el test de Turing.





Aun así, aunque el campo de la AI no es nuevo, en los últimos años se han realizado avances vertiginosos, con aplicaciones que van desde los coches de conducción autónoma hasta el diagnóstico médico, pasando por el *trading* automático, el reconocimiento facial, la gestión de la energía, la ciberseguridad, la robótica o la traducción automática, por citar algunas.

Una característica diferencial de la AI actual está precisamente ligada con la definición de McCarthy antes mencionada: no se limita a métodos observables, y, cuando alcanza cierto nivel de sofisticación, plantea problemas de interpretabilidad. En otras palabras: los modelos de AI tienden a tener una elevada tasa de acierto, muy superior a los algoritmos tradicionales; pero en cada caso concreto puede resultar extremadamente complejo explicar por qué el modelo ha producido un resultado determinado.

Aunque hay aplicaciones de la AI en las que no es tan relevante ser capaces de comprender o explicar por qué el algoritmo ha arrojado un valor concreto, en muchos casos resulta esencial y es un requerimiento regulatorio. Por ejemplo, en la Unión Europea, de acuerdo con el Reglamento General de Protección de Datos (GDPR), los consumidores tienen lo que se conoce como el "derecho a una explicación"[6]:

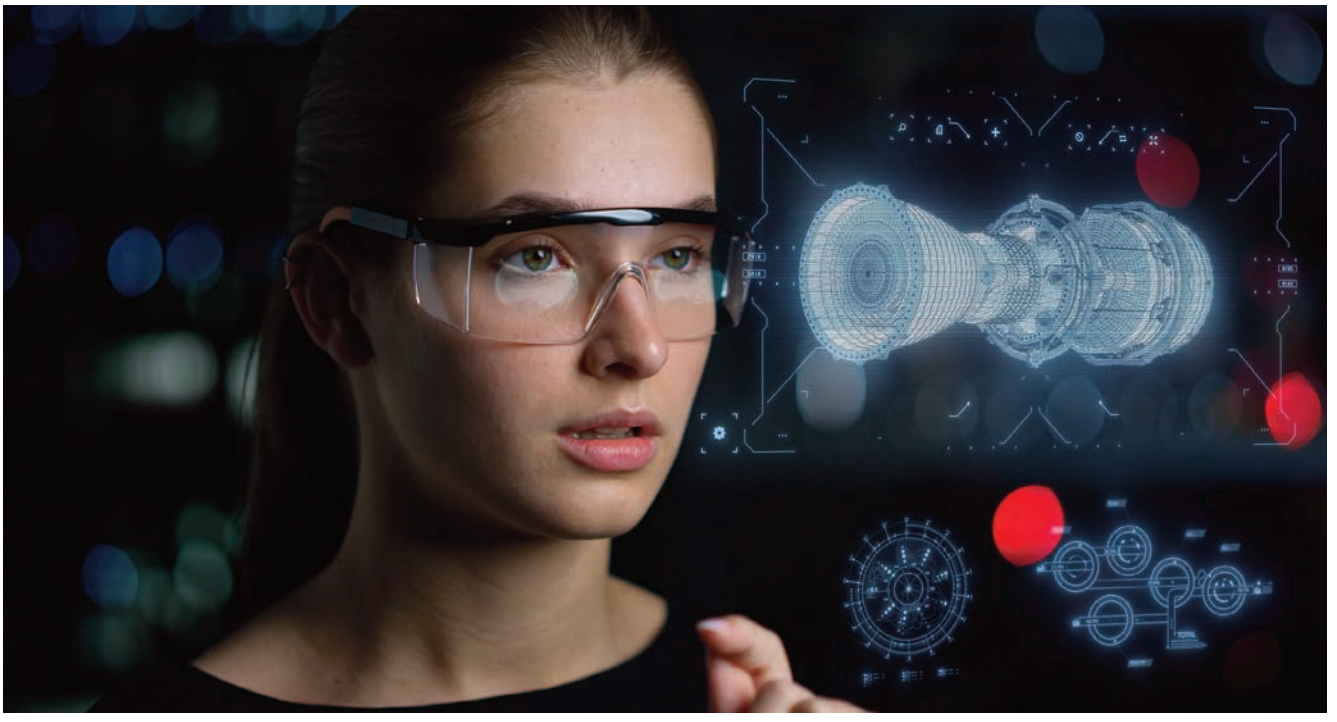
> [...] no ser objeto de una decisión [...] que se base únicamente en el tratamiento automatizado [...], como la denegación automática de una solicitud de crédito en línea, [...] [en la que] no medie intervención humana alguna", y tiene derecho "a recibir una explicación de la decisión tomada [...] y a impugnar la decisión".

Todo esto ha llevado al desarrollo de la disciplina de la inteligencia artificial explicable (XAI), que es el campo de estudio que pretende conseguir que los sistemas de AI resulten comprensibles para el ser humano[7], por contraposición a la noción de "caja negra" *(black box)*, que alude a los algoritmos en los que solo son observables los resultados y se desconoce el funcionamiento del modelo, o no se consigue explicar el fundamento por el cual se arrojan dichos resultados.

Se puede considerar[8] que un algoritmo se enmarca en la disciplina XAI si sigue tres principios: transparencia, interpretabilidad y explicabilidad. La transparencia se da si se pueden describir y justificar los procesos que calculan los parámetros del modelo y producen los resultados. La interpretabilidad describe la posibilidad de entender el modelo y presentar cómo toma decisiones de una manera comprensible para los humanos. La explicabilidad alude a la capacidad de descifrar por qué una determinada observación ha recibido un valor concreto. En la práctica, son tres términos muy ligados y con frecuencia se emplean de manera intercambiable, ante la falta de consenso sobre sus definiciones precisas[9].

Con el objetivo de conseguir dichos principios se pueden establecer básicamente dos estrategias de abordaje: o bien desarrollar algoritmos que son interpretables y explicables por

[6] GDPR (2018), Recital 71.
[7] Vilone et al. (2021). Doctora en Inteligencia Artificial, School of Computer Science, Technological University Dublin.
[8] Roscher et al. (2020). Científico de Datos de la Universidad Técnica de Múnich.
[9] Marcinkevics et al. (2020). Investigador en el Departamento de Informática, ETH Zürich.



su naturaleza (como las regresiones lineales, los modelos logísticos o multinomiales, o ciertos tipos de redes neuronales profundas, entre otros), o bien utilizar técnicas de interpretabilidad como herramientas para conseguir cumplir con estos principios[10].

La XAI, por tanto, se ocupa tanto de las técnicas para intentar explicar el comportamiento de determinados modelos opacos *(black box)* como del diseño de algoritmos inherentemente interpretables *(white box)*[11].

La XAI es fundamental en el desarrollo de la AI, y por tanto para los profesionales que trabajan en contacto con ella, por al menos tres factores:

- Contribuye a generar confianza en la toma de decisiones basadas en modelos de AI; sin esta confianza, los usuarios de estos modelos podrían mostrar resistencia a su adopción.
- Es un requerimiento regulatorio en determinados ámbitos (p. ej., protección de datos, protección del consumidor, igualdad de oportunidades en la contratación de empleados, regulación de modelos en banca).
- Favorece la mejora y el robustecimiento de los modelos de AI (p. ej., mediante la identificación y la eliminación de sesgos, la comprensión de la información relevante para producir un determinado resultado o la anticipación de posibles errores en observaciones no contempladas en la muestra de entrenamiento del modelo). Todo ello revierte en el desarrollo de algoritmos éticos, y permite focalizar los esfuerzos en las organizaciones en la identificación y aseguramiento de la calidad de los datos que son relevantes en los procesos de decisión.

Aunque el desarrollo de sistemas de XAI está recibiendo gran atención por parte de la comunidad académica, la industria y los reguladores, todavía plantea numerosos desafíos.

En este documento se repasan el contexto y los fundamentos de la XAI, incluyendo la normativa al respecto y sus implicaciones en la organización; el estado del arte y las principales técnicas de XAI; y los avances y retos sin resolver en la XAI. Por último, se proporciona un caso de estudio de XAI, para contribuir a ilustrar su aplicación práctica.

[10] iDanae (2022). Cátedra (inteligencia, datos, análisis y estrategia) en Big Data y Analytics, que surge gracias a la colaboración entre Management Solutions y la Universidad Politécnica de Madrid (UPM) en los campos formativo, científico y técnico, y tiene como objetivo promover la generación de conocimiento, difusión y transferencia de tecnología, y fomento de la I+D+i en el área de Análisis de Datos.

[11] Sudjianto et al. (2011). Responsable de Riesgo de Modelo de Wells Fargo.