

DANIEL GONZÁLEZ LAGIER

HECHOS Y MÁQUINAS

¿Puede la inteligencia artificial razonar
sobre la prueba?

Marcial Pons

MADRID | BARCELONA | BUENOS AIRES

2026

ÍNDICE

	Pág.
INTRODUCCIÓN.....	11
CAPÍTULO I. INTELIGENCIA ARTIFICIAL.....	15
1. ¿Cómo funciona la IA? Estrategias, métodos o «enfoques» para su desarrollo	15
2. Algunas precisiones sobre las capacidades de la IA	20
2.1. ¿Las máquinas entienden? El problema de la semántica..	20
2.2. ¿Tienen consciencia las máquinas?	23
2.3. ¿Las máquinas realmente aprenden «solas»?	26
2.4. El problema del sentido común	27
CAPÍTULO II. RAZONAMIENTO PROBATORIO: OTRA VUELTA DE TUERCA.....	33
1. El núcleo del razonamiento probatorio	33
2. La inferencia probatoria. Sus elementos	34
2.1. El modelo de Toulmin de representación de los argumen- tos.....	34
3. La inferencia probatoria desde el modelo de Toulmin.....	40
3.1. Pretensión: La hipótesis.....	40
3.2. Razones: Los datos o hechos probatorios.....	42

	<u>Pág.</u>
3.3. La garantía: generalizaciones empíricas, presunciones normativas y definiciones	42
3.3.1. Las generalizaciones empíricas.....	42
3.3.2. Tipos de generalizaciones empíricas.....	44
3.3.3. La «proyectabilidad» de las generalizaciones empíricas	48
3.3.4. Presunciones normativas y definiciones.....	50
3.4. El respaldo en las inferencias probatorias	51
4. Los criterios de corrección de la inferencia probatoria	54
 CAPÍTULO III. INTELIGENCIA ARTIFICIAL Y RAZONAMIENTO PROBATORIO. UNA CONVERSACIÓN CON CHATGPT	 57
1. Deducción, inducción y generalizaciones empíricas: límites formales y contextuales de la IA.....	58
2. Abducción e inferencia a la mejor explicación: un reto para la IA..	68
3. ¿El razonamiento de la IA es justificatorio?.....	75
4. La consciencia y el razonamiento probatorio de la IA.....	94
5. Emociones, razonamiento probatorio y límites de la IA.....	107
6. Sentido común, sesgos y razonamiento probatorio de la IA.....	117
7. Un resumen y dos cuestiones finales.....	129
 APÉNDICE I. AUTOEVALUACIÓN DE LA FIABILIDAD DE LAS RESPUESTAS DE CHATGPT Y LA CUESTIÓN DE LA AUTORÍA.....	 135
 APÉNDICE II. UN COMENTARIO SOBRE LAS VERSIONES POSTERIORES DE CHATGPT	 143
 BIBLIOGRAFÍA.....	 147

INTRODUCCIÓN

Ada Lovelace, de soltera Ada Byron (hija de Lord Byron y de una adinerada aristócrata, cuya relación con el poeta duró poco y acabó mal), es considerada una de las pioneras de la programación informática. Su madre, en un intento de exorcizar la «locura poética» que su hija pudiera haber heredado de su padre, le procuró una esmerada educación en matemáticas y ciencias, lo que le permitió trabajar con el matemático Charles Babbage. Ada Lovelace quedó asombrada por la «máquina analítica» ideada (pero nunca completamente construida) por este a comienzos del siglo XIX, para la que creó, junto con Babbage, uno de los primeros algoritmos informáticos conocidos. Sin embargo, a pesar de estimar que la capacidad del artefacto iba mucho más allá de realizar diversos cálculos con números, pensaba que las máquinas nunca podrían ser —a diferencia de los poetas— creativas (y, por tanto, no podrían ser genuinamente inteligentes), porque solo eran capaces de seguir las instrucciones previamente proporcionadas¹.

La opinión de Lovelace fue considerada por el propio Alan Turing como una de las principales objeciones a la posibilidad de una IA comparable a la humana². Una de las características de la inteligencia humana es la capacidad de enfrentarse satisfactoriamente a situaciones nuevas, cosa que la IA no podría hacer, según

¹ OLIVEIRA Y FIGUEIREDO, 2024: 4.

² TURING, 1984: 35 y ss.

esta crítica. Sin embargo, un grupo de ingenieros de Microsoft ha publicado no hace mucho un informe en el que aseguran que un modelo temprano de GPT-4 (que ya ha sido ampliamente superado) exhibía «destellos de inteligencia general artificial»³. La IA general (a diferencia de la específica), no se ocupa del razonamiento en una actividad determinada, sino que aspira a reproducir la inteligencia humana, con su capacidad de desenvolverse en distintos ámbitos y de resolver problemas de todo tipo. La IA general debería ser —al menos— capaz de superar el conocido test de Turing, es decir, debería ser capaz de interactuar con un ser humano sin que este pueda detectar que es una IA⁴. La conclusión del informe de Microsoft es la siguiente: «El argumento central de nuestro trabajo es que GPT-4 alcanza una forma de inteligencia general, mostrando destellos de inteligencia general artificial. Así lo demuestran sus capacidades mentales básicas (como el razonamiento, la creatividad y la deducción), el abanico de temas en los que ha adquirido experiencia (como la literatura, la medicina y la programación) y la variedad de tareas que es capaz de realizar (por ejemplo, jugar, utilizar herramientas o explicarse a sí mismo)». Y al final del informe se preguntan: «¿Cómo razona, planifica y crea? ¿Por qué exhibe una inteligencia tan general y flexible cuando en el fondo no es más que la combinación de componentes algorítmicos simples —descenso gradiente y transformadores a gran escala— con cantidades de datos extremadamente grandes? Estas preguntas forman parte del misterio y la fascinación de los LLM [grandes modelos de lenguaje]»⁵.

En este trabajo pretendo reflexionar sobre las dificultades que se plantean para la construcción de un «Juez IA» de los hechos del

³ Muchas opiniones vaticinan lo cerca que está la «singularidad» de la IA, esto es, el momento en que la IA alcanzará la inteligencia humana o la superará. Una enumeración de este tipo de posturas (tanto de quienes advierten del peligro de la singularidad como de quienes ven en ella una oportunidad de mejora de las condiciones humanas) puede verse en MADRID, 2024: 10 y ss.

⁴ TURING, 1984: 9 y ss. No obstante, hay razones para pensar que el test de Turing no es tan indicativo de una inteligencia general como pudiera parecer. Jack Copeland ha señalado las siguientes: excesiva adscripción al conductismo, excesiva atención a las capacidades conversacionales, posibilidad de falsos positivos (casos de simulación eficaz), posibilidad de falsos negativos (en casos de inteligencia sin capacidades lingüísticas) o posibilidad de programación específica para superar el test. COPELAND, 1996.

⁵ BUBECK *et al.*, 2023.

caso⁶. La tesis que trataré de sostener es que un «Juez IA» (que pueda juzgar autónomamente los hechos) debería tener habilidades propias de una IA general, porque el razonamiento probatorio tiene un nivel de dificultad tal que solo una IA general plenamente desarrollada podría acometerlo con la solvencia de los (buenos) operadores jurídicos humanos⁷. Para justificar esta tesis trataré de mostrar que el razonamiento probatorio remite a capacidades de los seres humanos (como la intuición, el sentido común, la consciencia, la comprensión de significados, etc.) que no están al alcance de IA débiles o específicas y que todavía no es claro que puedan implementarse en un futuro en una máquina para construir una IA fuerte (y que, desde luego, no poseen las IA actuales). No se trata de revisar los retos (tampoco las aportaciones, que sin duda tiene) que la IA podría plantear para una teoría del razonamiento probatorio sino, por el contrario, los que la complejidad de este último plantea a la IA y los límites (hoy por hoy) de esta última. No pretendo, por otra parte, negar la posibilidad futura de una IA general que supere satisfactoriamente estos límites —tal cosa quedaría fuera de mis conocimientos y competencias—, sino solo mostrar las dificultades que, como condiciones necesarias de una verdadera IA general, habría que resolver. Lo que hay en el transfondo de esta discusión es una cuestión debatida: ¿los presupuestos teóricos de la IA pueden ser una teoría adecuada de la mente (incluyendo la humana)? ¿La relación entre la mente y el cerebro es como la relación entre el *software* y el *hardware* de un ordenador o computadora? ¿El razonamiento, el pensamiento, puede «computarse», expresarse completamente en algoritmos, en un lenguaje formal, o «emerger» a partir de operaciones de este tipo? No tengo contestaciones firmes a estas preguntas (aunque sí intuiciones contrarias a la posibilidad cercana de una IA general), pero plantearlas a propósito del razonamiento probatorio puede servir al menos como una reflexión acerca de la naturaleza de este último y permitirnos comprenderlo mejor.

⁶ Tomo la denominación de BONSIGNORE, 2021. Para un análisis de los cambios que la IA puede inducir en el derecho en relación con las decisiones sobre cuestiones normativas, véase AGUILÓ, 2024.

⁷ Cosa distinta es el uso supervisado de IA para apoyar a los operadores jurídicos en sus decisiones, que no pretendo negar en absoluto.

Divido el trabajo en tres partes: en la primera, expongo brevemente algunas de las estrategias para el desarrollo de la IA, más que nada con la intención de explicar, aunque superficialmente, la terminología necesaria para entrar en estos temas, y a continuación señalo algunos límites de la IA que muchas veces quedan ocultos en nuestra manera de hablar sobre ella (como su carencia de capacidades semánticas). En la segunda, presento una forma de entender el razonamiento probatorio en el ámbito del proceso judicial. Se trata de una presentación que sigue de cerca lo que he dicho en otros muchos trabajos, pero profundizando («otra vuelta de tuerca») en algunos puntos (especialmente en lo relacionado con las generalizaciones empíricas). La tercera parte es, a la vez, la más y la menos original. Es la más original porque trato de profundizar en los desafíos que la complejidad del razonamiento probatorio plantea a la IA a través de una conversación con una IA (ChatGPT 4o⁸), en una suerte de «experimento argumentativo». Es la menos original porque está expuesta reproduciendo literalmente las preguntas (mías) y las respuestas (de la IA), en lugar de redactar de propia mano las conclusiones a las que se ha ido llegando. Creo que esta manera de actuar tiene una ventaja, que he querido preservar: se puede ver de primera mano y de manera ejemplificada tanto las capacidades (que son muchas) como los límites (sutiles, pero no menos importantes) de la manera de razonar de una IA. El lector menos interesado puede ir directamente a la síntesis (también redactada por ChatGPT) de esta parte. Por último, incluyo un «Apéndice», donde (también en forma de conversación con la IA) trato de adelantarme a las posibles dudas acerca de la fiabilidad de las contestaciones de la IA y a las consideraciones críticas de carácter ético por esta forma de presentar ideas y escribir parte de este trabajo.

⁸ En el momento en que este texto se publique, dada la velocidad a la que se suceden los avances en inteligencia artificial y los tiempos de edición habituales, es más que probable que existan versiones posteriores de ChatGPT u otros LLM más desarrollados con los que «tener» esta conversación, pero espero que esto no cambie en lo sustancial las conclusiones de este «experimento».

CAPÍTULO I

INTELIGENCIA ARTIFICIAL

1. ¿CÓMO FUNCIONA LA IA? ESTRATEGIAS, MÉTODOS O «ENFOQUES» PARA SU DESARROLLO

A la hora de desarrollar la IA se han adoptado dos tipos de estrategias: las estrategias «de arriba a abajo» (o enfoques simbólicos) y las estrategias «de abajo a arriba» (o enfoques conexionistas)¹. Las primeras consisten en la construcción de representaciones simbólicas del conocimiento (por ello se habla también de IA simbólica), que son procesadas de acuerdo con leyes o reglas de lógica matemática (también expresadas simbólicamente), que permiten extraer o concluir información a partir de ese conocimiento². Usando el lenguaje de la lógica, podríamos decir que operan fundamentalmente de manera «deductiva» y que la información que extraen en cierta manera ya estaba contenida en el conocimiento que se les ha aportado. La idea de fondo es que cualquier sistema físico puede tener comportamientos inteligentes si dispone de información y la ma-

¹ LÓPEZ DE MÁNTARAS y MESEGUER GONZÁLEZ, 2017. AMORUSO, BRUNO y DOMININO, 2007. Una revisión temprana de los presupuestos filosóficos y psicológicos de ambas estrategias puede verse en DREYFUS, 1972.

² BERLANGA, 2016: 9.

nipula de acuerdo con ciertas reglas previamente suministradas al sistema³. Desde un punto de vista filosófico, incluso se ha sostenido que la inteligencia humana consiste también en el procesamiento de símbolos (es la IA entendida como teoría de la mente). Este enfoque dio lugar a la elaboración de numerosos sistemas expertos, tecnologías «de una sola utilidad» que son adecuadas para aquellas actividades (como determinados juegos o el diagnóstico médico) que tienen un alcance cerrado y son fácilmente codificables. Un sistema experto requiere de la introducción de la información por parte de expertos (un conocimiento que debe estar libre de sesgos y de errores) y de un algoritmo (las reglas de procesamiento de la información) que permitan extraer la nueva información⁴. Estos sistemas pueden ser muy útiles, pero son claramente limitados si los comparamos con la inteligencia humana. Su éxito ha sido paradójico: ha resultado mucho más fácil construir sistemas capaces de realizar operaciones sumamente complejas para los seres humanos (como resolver complicados problemas matemáticos) que sistemas que sepan realizar tareas sencillas para nosotros, incluso para niños, como reconocer imágenes (distinguir un gato de un perro, por ejemplo) o entender el lenguaje natural. A esto se le conoce como la «paradoja de Moravec»⁵.

Las estrategias «de abajo a arriba» han desplazado en buena medida a las anteriores (aunque es más exacto decir que las han incorporado y superado). Consciente de las dificultades para construir una máquina inteligente, Alan Turing había propuesto la idea de una «máquina-niño» para que, a través de un proceso de aprendizaje, pueda adquirir por sí misma inteligencia y racionalidad⁶. Surge así la idea del aprendizaje automático, esto es, de construir sistemas que, a partir de varios ejemplos (el conjunto de datos de entrenamiento), sean capaces de establecer correlaciones o patrones entre ellos. Por ejemplo, si se le proporciona a la máquina la latitud y la temperatura media de varias ciudades, esta puede establecer (aunque sea aproximadamente, dado que hay otros factores que no han sido tenidos en cuenta) cuál es la relación entre latitud y temperatu-

³ NEWELL y SIMON, 1976.

⁴ BERLANGA, 2016: 9 y ss.

⁵ MORAVEC, 1988.

⁶ TURING, 1984: 30 y ss.