

INTELIGENCIA ARTIFICIAL Y RAZONAMIENTO JURÍDICO: UNA APROXIMACIÓN A SU ALCANCE Y LÍMITES*

Artificial Intelligence and Legal Reasoning: An Approach to Its Scope and Limits

GEMA MARÍA MARCILLA CÓRDOBA**

Fecha de recepción: 31/03/2025

Fecha de aceptación: 19/05/2025

Anales de la Cátedra Francisco Suárez

ISSN: 0008-7750, núm. 60 (2026), 325-353

<https://doi.org/10.30827/acfs.v60i.33405>

RESUMEN Este trabajo lleva a cabo una aproximación al alcance y los límites de la Inteligencia Artificial en el razonamiento jurídico, atendiendo a cómo esta replica las operaciones racionales que caracterizan a la interpretación y aplicación del Derecho, paradigmáticamente representada en el ámbito judicial. En primer lugar, se revisa el concepto legal y técnico de la IA. En segundo lugar, se analiza el debate filosófico sobre la posibilidad de reproducir las facultades cognitivas humanas en entes sintéticos. A continuación, se describe cómo funcionan los grandes modelos de lenguaje, capaces de simular razonamientos teóricos y prácticos. Finalmente, se plantean las posibilidades y limitaciones que, desde distintas concepciones del Derecho, tiene la IA en la argumentación jurídica, en especial en relación con la justificación de las decisiones judiciales.

Palabras clave: Inteligencia Artificial, teoría de la mente, grandes modelos de lenguaje, razonamiento jurídico.

ABSTRACT This paper examines the scope and limits of Artificial Intelligence in legal reasoning, exploring how AI replicates the rational processes underlying the interpretation and application of the law, paradigmatically represented within the judicial context. Firstly, the legal and technical concept of AI is reviewed. Secondly, the philosophical debate surrounding the possibility of reproducing human cognitive faculties in synthetic entities is analysed. Subsequently, the functioning of large language models, capable of simulating theoretical and practical reasoning, is described. Finally, the paper addresses the possibilities and limitations of AI in legal argumentation from different conceptions of law, particularly concerning the justification of judicial decisions.

Keywords: Artificial Intelligence, Theory of Mind, Large Language Models, Legal Reasoning.

* Para citar/citation: Marcilla Córdoba, G. M. (2026). Inteligencia artificial y razonamiento jurídico: una aproximación a su alcance y límites. *Anales de la Cátedra Francisco Suárez* 60, pp. 325-353.

** Universidad de Castilla-La Mancha. Facultad de Derecho. Plaza de la Universidad, 1, 02006 Albacete (España). Correo electrónico: gema.marcilla@uclm.es

1. INTRODUCCIÓN

Los avances en inteligencia artificial (IA) suponen la automatización de procesos hasta ahora propios del ser humano. En la práctica del Derecho surge el interrogante de hasta qué punto la IA tiene virtualidad en el razonamiento jurídico, en particular en la toma de decisiones judiciales. Este trabajo sostiene la hipótesis de que, con la adecuada implicación institucional, el auxilio de la IA es beneficioso para la práctica judicial, y no solo desde la perspectiva de la celeridad y la eficacia en la toma de decisiones, sino también desde la óptica de la calidad argumentativa de las sentencias, reforzando la legitimidad jurisdiccional.

Desde un enfoque interdisciplinar se partirá del concepto de IA, considerando su definición legal y técnica, para después abordar si es posible reproducir la inteligencia humana en sistemas artificiales. Tras repasar la operatividad de la IA, tanto en sus primeros momentos, a través de sistemas expertos, como en la actualidad, a través de redes neuronales y grandes modelos de lenguaje, se examina su potencial y límites en el razonamiento jurídico, a fin de dilucidar la utilidad de la IA en la práctica argumentativa de jueces y tribunales, y, de ser así, qué salvaguardas requeriría.

Ahora bien, no se trata únicamente de dilucidar qué puede hacer la IA en términos de eficiencia o de simulación cognitiva. Como advierte Solar Cayón (2025), la cuestión decisiva es si la inteligencia artificial jurídica contribuye a reforzar o, por el contrario, a debilitar el ideal del imperio de la ley. Esto exige desplazar la atención desde las definiciones técnicas o legales hacia un análisis más amplio: cómo incide la IA en la transparencia de la creación normativa, en la cognoscibilidad de las reglas jurídicas y en la imparcialidad de quienes aplican el Derecho. En este marco, los retos que plantea la IA al razonamiento jurídico deben medirse en función de su capacidad para salvaguardar los principios estructurales del Estado de Derecho.

2. CONCEPTO LEGAL Y TÉCNICO DE INTELIGENCIA ARTIFICIAL

El Reglamento (UE) 2024/1689, que trata de proporcionar un marco jurídico uniforme para desarrollar y usar los sistemas de IA en la Unión Europea (art. 3), destaca en la definición la autonomía y la capacidad de adaptación de la IA, así como la posibilidad de generar decisiones o recomendaciones que pueden influir en entornos físicos o virtuales.

En su Considerando 12, el Reglamento subraya la complejidad de definir la IA (Cotino Hueso, 2024, p. 115; Huergo Lora & Díaz González,

2024), señalando la necesidad de una definición que distinga claramente la IA de la informática tradicional, pero que resulte flexible, es decir, comprensiva de los distintos modelos de IA, capaz de sobrevivir al constante avance técnico y armonizable con estándares de organismos como la OCDE¹.

La crítica definición del Reglamento de IA contrasta con una de las definiciones técnicas más influyentes en la historia de la IA, procedente del documento titulado “A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence” (McCarthy *et al.*, 1955, p. 11). Detrás de este proyecto se hallaba un grupo de investigadores: John McCarthy, Marvin Minsky, Nathaniel Rochester y Claude Shannon, que se unieron para explorar la posibilidad de que las máquinas pudieran emular comportamientos “que denominamos inteligentes” cuando son llevados a cabo por los seres humanos. El espíritu interdisciplinar de estos científicos, provenientes de ámbitos como la lógica, las matemáticas y la computación, impregnó el proyecto de un entusiasmo pionero que sentó las bases de la IA como disciplina. Su propuesta describía la inteligencia artificial como la “ciencia y la ingeniería de crear máquinas inteligentes”, orientada a conseguir que los dispositivos informáticos reprodujeran las capacidades cognitivas consideradas propias de la mente humana.

La IA actual difiere de la informática tradicional en la naturaleza de sus algoritmos, que aprenden de los datos (Machine Learning, ML) y “se reprograman” tras identificar patrones. La informática tradicional automatiza la ejecución de tareas, mientras que la IA actual automatiza el aprendizaje. Esto se ha potenciado con la tecnología de redes neuronales artificiales (Deep Learning, DL), haciendo que la IA alcance predicciones cada vez más precisas (Goodfellow *et al.*, 2016).

Un punto clave es la capacidad de la IA para procesar enormes cantidades de datos. A diferencia de un modelo lineal tradicional (p. ej., para predecir el comportamiento de la bolsa con variables fijas), la IA aprende de múltiples variables, incluso si no eran inicialmente identificadas por humanos. Su evolución ha pasado por varias etapas: desde la IA simbólica

1. Precisamente por la complejidad de la definición, y en cumplimiento del artículo 96.1. f) del Reglamento, la Comisión Europea ha tratado de aclarar el art. 3, tras las aclaraciones sobre las prácticas prohibidas de inteligencia artificial. <https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application>.

de mediados del siglo xx hasta la eclosión del aprendizaje profundo y la IA generativa en 2023².

3. LA INTELIGENCIA ARTIFICIAL EN LA TEORÍA DE LA MENTE

El objetivo de la IA es llevar a cabo actividades cognitivas humanas, como el aprendizaje y la resolución de problemas. Por ello, con la IA se

-
2. La consciencia de ese carácter críptico de la definición, inevitable por otra parte al intentar dar flexibilidad a la misma en lo que respecta a la evolución tecnológica, la Comisión Europea ha desgranado ese concepto de IA dividido en siete elementos: Primero, es un *sistema basado en una máquina*. Es decir, nace y opera sobre hardware y software. Segundo, está *diseñado para funcionar con distintos niveles de autonomía*, es decir, independiente de la actuación humana en algún grado: se excluye el funcionamiento puramente manual, pero se admite todo lo que suponga generación de resultados sin control humano directo en cada paso. Tercero, *puede mostrar capacidad de adaptación tras el despliegue*, lo que equivale a autoaprendizaje. Cuarto, todo sistema *actúa orientado por objetivos*, que pueden ser *explícitos*, es decir, codificados por el desarrollador, o *implícitos*, esto es, deducidos del entrenamiento o de la interacción con el entorno. Conviene distinguirlos de la “finalidad prevista”, que es externa y viene dada por el uso para el que el proveedor concibe el sistema. Quinto, *la característica definitoria es la inferencia*: el sistema infiere, a partir de la entrada, la manera de generar la salida. Aquí se marca la frontera con el software tradicional. La Comisión aclara que la inferencia opera en dos planos del ciclo de vida. Durante la “construcción” se emplean técnicas que habilitan esa capacidad —aprendizaje automático en sus variantes supervisada, no supervisada, autosupervisada, por refuerzo y aprendizaje profundo; y estrategias lógicas o basadas en conocimiento, como sistemas expertos, ontologías o motores de deducción—. En la “utilización”, esa capacidad se traduce en resultados con efectos en el entorno. Sexto, *las salidas del sistema adoptan cuatro formas típicas: predicciones, contenidos, recomendaciones y decisiones*. Las predicciones estiman valores desconocidos en contextos complejos y dinámicos. Los contenidos remiten a la IA generativa en cualquiera de sus modalidades. Las recomendaciones sugieren acciones o selecciones, y pueden convertirse en decisiones si se ejecutan automáticamente. Las decisiones, por su parte, automatizan procesos tradicionalmente reservados al juicio humano. Séptimo, *esas salidas pueden influir en entornos físicos o virtuales*. En otros términos, la IA no es pasiva, sino que afecta a objetos, flujos de datos y ecosistemas digitales, y es esta interacción con el entorno lo fundamental. Además, la definición distingue entre *fase de construcción y fase de utilización*. No exige que los siete elementos estén presentes de manera continua en ambas. Algunos aparecerán en el desarrollo del sistema y otros en su despliegue. Por otra parte, la definición excluye los sistemas de optimización matemática clásica, los de aceleración de métodos basados en el “tratamiento básico de datos”, el software de gestión o visualización que ejecuta reglas fijas, sin aprendizaje, los sistemas que resuelven por reglas prefijadas y los predictores que ofrecen promedios o referencias simples sin capturar patrones complejos. En suma, la frontera entre lo que es IA y lo que no, la traza *la capacidad real de inferir y de generar salidas que inciden en el entorno*. Pueden afectar a un entorno físico (pensemos en un coche autónomo que frena al detectar un obstáculo), o a un entorno virtual (pensemos en la recomendación de una película por una plataforma).

plantea el debate sobre si la mente humana puede ser replicada por entes sintéticos.

3.1. La postura no materialista: imposibilidad de replicar la mente

En este debate son varios los filósofos y científicos que expresa su negativa en relación a que la mente humana pueda ser reproducida en sistemas artificiales. Para John Searle (1980, pp. 417-422; 1992, pp. 58-60), por ejemplo, la experiencia consciente y la intencionalidad trascienden la mera computación: su experimento de la “habitación china” ilustra cómo un programa informático puede manipular símbolos, lo cual no significa que comprenda su significado.

Thomas Nagel (1974, p. 436; 2012, p. 44) recurre al concepto de “cualia” para resaltar el carácter subjetivo de la vivencia consciente, alegando que no es posible reproducirla artificialmente por cuanto no es algo que pueda ser explicado a través de descripciones objetivas.

Roger Penrose (1989) afirma que la conciencia responde a procesos cuánticos no computables, mientras que Hubert Dreyfus (1992) subraya la imposibilidad de separar la inteligencia humana de la corporalidad y el entorno en que se desarrolla. Por su parte, Luciano Floridi (2024, pp. 22-23; 120-128) advierte que la IA disocia agencia e inteligencia: representa una forma de agencia que puede resolver problemas con eficacia, pero sin la comprensión auténtica propia de la mente humana. En conjunto, estas posturas señalan que la IA, tal como la conocemos, no alcanzaría la experiencia subjetiva ni la profundidad cognitiva que caracteriza a la inteligencia de los seres humanos.

3.2. La postura materialista: posibilidad de emular (o superar) la IA

Frente a las posiciones anteriores se encuentran los autores que adoptan una visión materialista de la mente. Sostienen que es cuestión de desarrollo técnico que puedan replicarse en soportes sintéticos las funciones cognitivas humanas. Por ejemplo, Hilary Putnam (1967, p. 45) y Daniel Dennett (1991, p. 177) consideran que toda la actividad mental puede ser descrita en términos de funciones de un sistema computacional.

Para el *funcionalismo* de Putnam, lo esencial no es la base material, biológica o artificial, en la que se desarrollan las funciones cognitivas. Lo importante es la función en sí. Por ello, si logramos reproducir las mismas

relaciones causales y transformaciones de la información que se dan en la mente, surgiría inteligencia.

Dennett (1991, p. 183) consolida este enfoque al proponer el *computacionalismo*, es decir, la tesis de que los procesos mentales pueden comprenderse como operaciones con símbolos y reglas, análogas al tratamiento informático de datos.

Más allá va David Chalmers (2010, pp. 35, 44, 61) [1996], quien plantea la posibilidad de que haya sistemas suficientemente complejos capaces de alcanzar experiencias subjetivas y estados conscientes. A su juicio, lo decisivo sería la estructura y la organización funcional, sin importar si está implementada en un cerebro biológico o en un dispositivo artificial.

Desde una perspectiva diferente a la de los autores previamente mencionados, Nick Bostrom (2016) presenta la idea de que la inteligencia artificial podría alcanzar un nivel de Superinteligencia, esto es, un tipo de cognición radicalmente distinta y, a la vez, muy superior a la humana. Para Bostrom, la IA atraviesa tres posibles estadios: la IA específica, focalizada en tareas concretas con gran eficiencia; la IA general, que abarcaría un amplio espectro de capacidades análogas a las humanas; y, finalmente, la Superinteligencia, capaz de superar de manera abrumadora nuestras facultades intelectuales. Dado que estos sistemas podrían automejorarse sin cesar, Bostrom alerta sobre la posibilidad de una “explosión de inteligencia” fuera de nuestro control.

En sintonía con esta visión, Tobias Rees plantea que la IA no necesita imitar metódicamente la estructura biológica de la mente humana para alcanzar u ofrecer formas de inteligencia, seguramente nuevas. Rees considera que la IA replantea nuestras nociones sobre la mente, la subjetividad y el conocimiento, considerando a las grandes empresas tecnológicas como auténticos “espacios filosóficos” donde se redefine la realidad contemporánea. Según el autor (Rees, 2019; 2022; 2025), los grandes modelos de lenguaje están contribuyendo a una verdadera “ruptura filosófica”, de manera que cada vez resulta menos pertinente comparar la IA con la inteligencia humana. Podría suceder incluso que la IA despliegue métodos de conocimiento y categorías conceptuales incomprensibles desde los esquemas humanos.

En síntesis, desde la corriente fisicalista o materialista, se rechaza la idea de un abismo insalvable entre la mente biológica y la máquina. Bajo este enfoque, es plausible reproducir o incluso superar las capacidades intelectuales humanas siempre que se alcance un desarrollo tecnológico capaz de emular las complejas funciones de la mente. Además, la propia evolución de la IA podría dar lugar a lógicas y estructuras cognitivas tan diferentes de las nuestras que perdería sentido equipararlas con la inteligencia humana,

inaugurando así nuevos horizontes para la creación y el entendimiento del conocimiento.

4. EL FUNCIONAMIENTO DE LA IA

Al margen del debate sobre la replicación total de la inteligencia humana, la tecnología actual únicamente alcanza la inteligencia artificial *específica*, orientada a tareas concretas. Considerando el objetivo central del trabajo, que es determinar el impacto de la IA en el razonamiento jurídico, es conveniente abordar tres puntos fundamentales: primero, el funcionamiento general de la IA (epígrafe 4); segundo, las tareas racionales simuladas por la IA en los grandes modelos de lenguaje (epígrafes 5 y 6); y tercero, la aplicación de esas tareas al ámbito jurídico (epígrafe 7).

4.1. La IA lineal y los sistemas expertos

En filosofía lógica y científica, las operaciones racionales básicas son la deducción y la inducción. La deducción consiste en obtener conclusiones necesarias a partir de premisas ciertas, siendo las conclusiones irrefutables si las reglas se aplican correctamente. Este método de razonamiento prevaleció inicialmente en la IA simbólica o lineal, característica de los sistemas expertos desarrollados en las décadas de 1970 y 1980. Un sistema experto resuelve problemas mediante reglas formales previamente definidas en lenguaje lógico o simbólico. Por ejemplo, en medicina, un sistema experto diagnosticará gripe si detecta síntomas como fiebre alta, tos seca y dolor muscular, aplicando estrictamente reglas preestablecidas (Newell & Simon, 1972).

Sin embargo, estos sistemas adolecían de limitaciones importantes derivadas de su naturaleza lógico-deductiva. Fuera de casos explícitamente previstos en su base de conocimiento, o ante contradicciones en las reglas, generan resultados incongruentes o se bloquean al no encontrar soluciones adecuadas (Gardner, 1987; Jackson, 1990). Por ello, el razonamiento deductivo en IA es limitado y poco eficaz ante problemas complejos con múltiples variables, como es habitual en el ámbito jurídico, y veremos más adelante.

4.2. La IA conexionista y el aprendizaje inductivo-estadístico

La evolución de la inteligencia artificial desde modelos simbólicos o lógicos hacia modelos conexionistas o de redes neuronales, implicó el paso

del uso de reglas predefinidas, al aprendizaje inductivo mediante el análisis estadístico de grandes conjuntos de datos. Mediante la estadística, la IA extrae patrones, es decir, regularidades observables en el análisis empírico de los datos (Domingos, 2015).

En efecto, la IA actual, especialmente en modalidades de aprendizaje automático (ML y DL), depende en gran medida de métodos estadísticos (Goodfellow, Bengio & Courville, 2016). Estos métodos permiten extraer patrones para mejorar algoritmos predictivos, como ocurre, por ejemplo, cuando plataformas digitales anticipan preferencias basadas en elecciones previas de usuarios. Además, los modelos se ajustan continuamente comparando predicciones con resultados reales (Freedman *et al.*, 2007).

Especial relevancia tienen para el análisis del razonamiento jurídico —que se verá a continuación— los modelos de lenguaje de gran escala (LLMs), que aprenden a través del análisis de amplios corpus textuales patrones estadísticos lingüísticos, es decir, gramaticales, semánticos y de uso, mediante técnicas como la tokenización e incrustación; naturalmente, los resultados que ofrecen estos modelos no son fruto de una comprensión semántica genuina, sino de previas asociaciones de expresiones y conceptos que se repiten recurrentemente en los datos de entrenamiento (Soler, Labeau & Clavel, 2024; Floridi & Chiriatti, 2020; Bender & Koller, 2020).

Resumiendo, la diferencia entre el funcionamiento de la IA lineal y la IA neuronal estriba en lo siguiente: la primera se aplica en sistemas expertos, que deducen conclusiones estrictamente lógicas desde reglas fijas y hasta el punto de que la base de conocimiento lo permite. La segunda se aplica al análisis de datos, como, por ejemplo, datos lingüísticos, obteniendo en muy poco tiempo y con gran exactitud patrones o regularidades, que, además, continuamente se están actualizando ante la introducción de nuevos datos y la mejora de los sistemas de análisis. La IA, gracias a su aprendizaje estadístico, es capaz, sin comprensión genuina, de simular distintos tipos de razonamiento humano, tales como la deducción lógica y el cálculo matemático, la inducción de leyes generales, el establecimiento de hipótesis explicativas a partir de datos incompletos o analogías. Tal simulación se lleva a cabo con una creciente calidad, como sucede con los denominados “agentes”, la última generación de LLMs, dotados de memoria y de método de pensamiento paso a paso o *chain of thought* (Calvo, 2024). De todos modos, hay más estudios de IA “razonadora” distintos de los LLMs, en particular la llamada “programación probabilística” de Russell (2015). Veamos la diferencia: los LLMs aprenden mediante asociaciones estadísticas implícitas a partir de los textos con los que se entrenan, codificando estas relaciones en vectores numéricos, llamados *embeddings*. Por ejemplo, ante la pregunta “¿qué animal maúlla?”, responden “el gato”, porque detectan esta conexión recurrente

(patrón), pero sin comprenderla realmente. En cambio, la programación probabilística que defiende Russell, *combina lógica y probabilidad para representar explícitamente conceptos, relaciones causales e incertidumbre*. En este enfoque, el conocimiento se expresa mediante reglas claras como: “Si es un gato, entonces probablemente maulla”. Al introducir como input (“se oye un maullido”), el modelo aplica la inferencia probabilística para llegar a la conclusión de que es altamente probable que se trate de un gato. Así, según Russell, esta unión de lógica y probabilidad permite “razonar de manera transparente y fundamentada sobre conceptos y relaciones causales en condiciones de incertidumbre” (Russell, 2015, p. 88), algo imposible con el aprendizaje meramente estadístico de los LLM.

5. SIMULACIÓN DE RAZONAMIENTO DEDUCTIVO, INDUCTIVO, ABDUCTIVO Y ANALÓGICO POR LOS LLMs

Los modelos de lenguaje de gran tamaño (LLMs) resultan muy relevantes para indagar sobre cómo la IA simula el razonamiento humano, ya sea *deductivo*, *inductivo*, *abductivo* o *analógico*. Conviene, pues, indicar las posibilidades de tal imitación y también las limitaciones de la IA, en comparación con las mismas capacidades humanas.

En efecto, la IA generativa basada en los LLMs, muestra aptitud para realizar, en primer lugar, operaciones de *razonamiento deductivo* o demostrativo, es decir, para obtener, a partir de unas premisas indiscutibles, una conclusión necesaria. Los LLMs resuelven silogismos o problemas lógico-formales, sobre todo cuando se les indica “pensar paso a paso” (técnica *chain-of-thought*) (Saparov & He, 2023). Por ejemplo, incluimos las siguientes dos premisas a modo de prompt en el chatbot de un LLM: “Todos los contratos requieren consentimiento. Este acuerdo es un contrato. Llega a una conclusión lógica”. El LLM responderá algo semejante a lo siguiente: “Basándonos en las premisas proporcionadas, podemos llegar a la siguiente conclusión lógica: 1. Todos los contratos requieren consentimiento; 2. Este acuerdo es un contrato. Por lo tanto, este acuerdo requiere consentimiento”. El LLM ha aprendido que la primera premisa establece una regla general; la segunda, un caso específico; y que, al aplicar la regla general (premisa 1) al caso específico (premisa 2), llegamos lógicamente o deductivamente a la conclusión.

Sin embargo, el rigor lógico de los LLMs es imperfecto, porque, como se ha señalado al explicar su funcionamiento y modo de aprender, carecen de una base de conocimiento formal. Su “conocimiento” es el resultado de

almacenar de forma implícita, relaciones lógicas (y matemáticas) al entrenarse. *Cuando se les presenta un problema lógico, reconocen la estructura de este sobre la base de patrones similares vistos en su entrenamiento*; identifican los elementos clave (premisas, conclusión) y sus relaciones; generan una respuesta que sigue la estructura lógica esperada, basándose en cómo se han formulado conclusiones similares en los datos de entrenamiento. Pero sus deducciones son imitativas y no analíticas (Dasgupta *et al.*, 2023; Eisape *et al.*, 2023).

Asimismo, los LLMs simulan el *razonamiento inductivo*. Este implica *generalizar, es decir, establecer reglas a partir de la observación de casos particulares*. Como señalamos, el aprendizaje de los LLMs es inductivo-estadístico, pues detectan patrones y los extrapolan a contextos similares (Qiu *et al.*, 2024). Y debido a esta posibilidad de hallar patrones en datos dados, pueden imitar mejor que ningún razonamiento, el razonamiento inductivo. Pongamos un ejemplo: suministramos a un LLM las estadísticas del INE de los delitos en España en 2025 y se le pregunta por los delitos que predominarán en 2025 bajo circunstancias similares. El LLM hará una predicción sobre la base de los patrones estadísticos previamente analizados.

Los LLMs se equivocan a menudo al aplicar reglas generales a escenarios nuevos o distintos, sobre todo en razonamientos *no monotónicos*, donde al añadir premisas a la premisa inicial, puede quedar debilitada la conclusión inicial. El que los LLMs no tomen en cuenta hacia qué conclusión apuntan las nuevas premisas ocurre porque, al operar con patrones aprendidos, puede que no hayan captado el patrón que subyace a la premisa añadida, o su relación conflictiva con los patrones propios de las premisas iniciales. Es decir, al recibir premisas adicionales que contradicen la conclusión previa, salvo que hayan aprendido tal patrón, carecen, a diferencia de los humanos, de la verdadera capacidad de revisar la inferencia original. Por ejemplo, si un LLM aprende que “en general, los antibióticos curan infecciones” y no ha asimilado “las infecciones virales no se curan con antibióticos”, no integrará adecuadamente la excepción en la conclusión (Han *et al.*, 2024). En cambio, si ha captado ambos patrones y su dimensión conflictiva, sí podrá aplicar la salvedad al caso particular (Qiu *et al.*, 2024).

Así, supongamos que se proporciona a un gran modelo de lenguaje (LLM) el dato de que en el año 2024 se denunció un porcentaje x de delitos de robo con violencia e intimidación. A continuación, se le pregunta cuál será el porcentaje aproximado de denuncias de tal delito en 2025, informándole de una cuestión más: un buen porcentaje de casos investigados de tal delito se basaron en denuncias falsas y, en respuesta a ello, se ha implantado un programa de IA que ayudará a la policía a discriminar denuncias falsas con mayor exactitud. Ante esta información adicional, el LLM probable-

mente señalará que la frecuencia de la denuncia de robo con violencia o intimidación disminuirá.

Es decir, cuanto más aprendan los modelos de lenguaje, más capaces serán de ofrecer respuestas coherentes y útiles a partir de nueva información proporcionada y de simular eficazmente el razonamiento inductivo. Por supuesto, no hay comprensión real de los fenómenos subyacentes, ni auténtico razonamiento causal, sino una extrapolación estadística basada en patrones lingüísticos previos (Bender y Koller, 2020).

En tercer lugar, los modelos de lenguaje demuestran capacidad en el *razonamiento abductivo*, es decir, el planteamiento creativo de una hipótesis explicativa de un estado de cosas en el que faltan datos que lo expliquen. Reichenbach (1938) denomina este momento creativo del razonamiento “contexto de descubrimiento”, en oposición al “contexto de justificación”, caracterizado por la demostración lógica o la verificación empírica de la hipótesis. La filosofía de la ciencia ha resaltado la importancia del razonamiento abductivo en la generación de conocimiento científico. Popper (1962)[1934] describió el método científico como un proceso que comienza con la formulación *creativa* de hipótesis, seguida de la deducción de consecuencias y su posterior contraste empírico, correspondiendo ese momento inicial al razonamiento abductivo. La abducción es, en resumen, un tipo de razonamiento que consiste en formular una hipótesis plausible para explicar un fenómeno, a través de un conjunto de datos o de circunstancias incompletas, verificándose posteriormente tal explicación tentativa (Peirce, 1976; Aliseda, 2006).

Los modelos de lenguaje exhiben capacidad abductiva, produciendo explicaciones viables y coherentes para escenarios incompletos (Dasgupta *et al.*, 2023). En este sentido, podríamos pedir a un LLM con entrenamiento jurídico que, a sabiendas de que una persona ha constituido una sociedad pantalla, sin actividad económica real, sin empleados ni instalaciones, que formule una explicación en términos delictivos. Aplicando razonamiento abductivo, el modelo podría plantear como hipótesis plausible que la sociedad se está utilizando para llevar a cabo un delito de blanqueo de capitales.

La simulación del razonamiento abductivo por los LLMs es posible porque captan correlaciones estadísticas (por ejemplo, creación de una sociedad pantalla y su relación con el delito de blanqueo de capitales) derivadas de su entrenamiento. Naturalmente, se trata, como en todos los tipos de razonamiento vistos, de una simulación pues los modelos carecen de “auténtica comprensión” de las relaciones causales implicadas en la hipótesis. Además, la validación empírica de la explicación propuesta debe hacerse externamente mediante evidencia adicional, que, evidentemente, el modelo no puede hacer autónomamente (Floridi y Chiriatti, 2020).

En este sentido, advierte González Lagier (2024, pp. 45-46), que en el ámbito del razonamiento probatorio la abducción constituye un límite estructural de la inteligencia artificial, pues esta solo produce correlaciones plausibles a partir de datos, *pero no explicaciones sustentadas en criterios epistémicos y normativos*. De ahí que la simulación abductiva de los modelos de lenguaje, aunque útil como recurso auxiliar, no pueda confundirse con el ejercicio genuino de formulación y justificación de hipótesis.

Por último, puede hablarse de LLMs y *razonamiento analógico*. Buscar patrones de relación entre ámbitos distintos es una facultad cognitiva elevada, y los LLMs más avanzados comienzan a mostrarla (Webb, Holyoak & Lu, 2023). Estos pueden inducir patrones abstractos y resolver analogías sin ejemplos previos. Esta competencia surge, como las demás, de su entrenamiento masivo en texto, que le brinda habilidades parecidas al razonamiento humano para transferir conocimiento de un contexto a otro (Webb *et al.*, 2023). Por supuesto que no comprenden las similitudes entre los supuestos, sino que detectan correlaciones en los datos que analizan (Dasgupta *et al.*, 2023). Es decir, identifican relaciones recurrentes y resuelven analogías reconociendo similitudes estructurales sin una comprensión profunda (Webb *et al.*, 2023).

Por ello, simulan cada vez mejor este tipo de razonamiento y abordan tareas antes consideradas exclusivamente humanas. Con todo, su fiabilidad varía según la complejidad y novedad del problema (Webb *et al.*, 2023). Podríamos por ejemplo preguntar a la IA por manifestaciones de la prohibición que establece el artículo 5.1. c) del Reglamento de Inteligencia Artificial, sobre el uso de sistemas de IA que evalúen o clasifiquen a personas físicas o grupos en función de su comportamiento social o características personales, lo que se conoce como “puntuación social”, y esta podría responder que se dan tales características en los sistemas de puntuación de los repartidores a domicilio que trabajan en plataformas web: aunque sus sistemas de puntuación están orientados a evaluar la actividad laboral, pueden tener efectos similares a la puntuación social si determinan el acceso del trabajador a futuras oportunidades laborales o incluso su exclusión de la plataforma.

Para concluir, puede decirse que, en general, los LLMs han demostrado capacidades similares a los cuatro tipos de razonamiento humano mencionados: deducen, inducen, ofrecen explicaciones abductivas y destacan en el razonamiento analógico. Sin embargo, no razonan como las personas, sino que son una especie de “parlanchines o loros estadísticos” que replican patrones observados en sus datos de entrenamiento. Por lo demás, a veces incorporan información irrelevante o falsa (“alucinaciones”) y esta falta de fiabilidad supone un importante defecto (Collins *et al.*, 2022).

En definitiva, en el estado tecnológico actual, la IA es específica, esto es, se limita a tareas concretas, que resuelve con gran exactitud y rapidez. Lo hace basándose en análisis estadístico mediante redes neuronales, gracias al cual puede ajustar el modelo y hallar la solución óptima para un problema a partir de múltiples patrones. Desde esta óptica, no existe “entendimiento”, “comprensión” ni “creatividad” en un sentido estricto, y la verificación empírica no estadística queda fuera de su alcance. Cuánta relevancia práctica tengan la representación mental de los conceptos y la creatividad—inexistentes en la IA— en la generación de conocimiento es otra cuestión.

6. SIMULACIÓN DEL RAZONAMIENTO PRÁCTICO POR LOS LLMs

La filosofía práctica sostiene que la racionalidad no se agota en “conocer” el mundo mediante la deducción, inducción o abducción, sino que abarca la capacidad de “orientar la acción” a fines moralmente pertinentes. Ser racional no consiste únicamente en conocer la realidad extrayendo conclusiones por métodos lógicos y empíricos, sino que implica articular juicios que guíen la conducta. Veamos si los LLMs pueden simular también este tipo de razonamiento.

6.1. *La relevancia de la experiencia subjetiva en el juicio moral y los límites de la IA*

¿Qué hace posible el juicio moral? Algunas teorías sostienen que, de manera inequívoca, las emociones y las cualidades como la empatía son fundamentales para adoptar criterios morales. Este es el caso de teorías como el Sentimentalismo Moral, la Teoría de la Moralidad basada en la Simpatía y la Ética del Cuidado.

El Sentimentalismo Moral, de David Hume y Adam Smith, sostiene que los juicios morales se fundamentan en sentimientos. Hume argumenta que las emociones de aprobación o desaprobación son esenciales para distinguir entre lo correcto y lo incorrecto (Hume, 2005 [1739-1740]; Radcliffe, 2018). Y Adam Smith añade que la capacidad de identificarse con los sentimientos de otro es crucial para el juicio moral (Smith, 1997, [1759]).

Geoffrey James Warnock defiende que el propósito de la moralidad es superar la “dificultad básica”, el egoísmo o la indiferencia, mediante la simpatía hacia los intereses ajenos. Las emociones son fundamentales para generar disposiciones virtuosas como la beneficencia y la imparcialidad (Warnock, 1971).

Asimismo, la ética del cuidado de Carol Gilligan sostiene que emociones como la empatía y el cuidado son centrales en la deliberación moral, por lo que el razonamiento ético no debe centrarse exclusivamente en principios abstractos, sino en relaciones y necesidades concretas de los demás (Gilligan, 1982).

En estas teorías, las emociones son el núcleo del razonamiento moral. Esta importancia de la emoción (también de la conciencia, la experiencia subjetiva y la intencionalidad) supone que, sin lugar a duda, en la actualidad, con una IA específica y no general, es imposible que esta desarrolle el juicio moral. Es más, si las teorías de la mente *no fisicalistas* están en lo cierto, *jamás será posible la automatización del razonamiento moral*. Los LLMs serían como loros repitiendo aserciones con contenido moral.

6.2. *El juicio moral dotado de objetividad y las posibilidades de la IA en su simulación*

Otras teorías éticas, por el contrario, estiman que la emoción y la subjetividad, inevitablemente presentes en las decisiones morales, son elementos que restan imparcialidad y objetividad. Por ello, apelan a la posibilidad de realizar juicios morales usando criterios objetivos. En este sentido pueden destacarse enfoques consecuencialistas y procedimentales de inspiración deontológica, que, tal vez, pudieran ser reproducidos por la IA.

A) Teorías morales consecuencialistas: la IA y el cálculo estadístico del bienestar general

El Utilitarismo evalúa las acciones según sus consecuencias, en términos de maximización del bienestar o la felicidad general. La utilidad puede ser calculada (Mill, 1984 [1863]). De manera semejante, el Análisis económico del Derecho establece la eficiencia económica o el uso racional de los recursos como criterio fundamental de decisión política y jurídica (Posner, 1972).

Pues bien, qué duda cabe de que los modelos de IA son capaces de tomar decisiones moralmente adecuadas en las claves del Utilitarismo y del Análisis Económico del Derecho, pues la resolución de problemas tiene criterios claros que se pueden fundar en el análisis y el cálculo estadístico de datos sobre la realidad. Es decir, la IA tiene el potencial de procesar enormes cantidades de datos y encontrar patrones mediante los que predecir las consecuencias de decisiones, identificando las que maximizan el

bienestar general o la eficiencia económica. Naturalmente, un problema moral previo que no podría responder la IA genuinamente es el relativo a qué deba entenderse por bienestar general, o si tiene o no algún límite la eficiencia económica; problemas estos a los que responden mejor las teorías éticas basadas en reglas procedimentales que conducen a consensos morales.

B) Teorías morales procedimentales de inspiración deontológica:
la *Machine Ethics*

El constructivismo de John Rawls y la ética discursiva de Jürgen Habermas son éticas procedimentales capaces de establecer los principios de justicia “objetivos”, en el sentido de que serían consensuados por seres racionales.

Según Rawls, los principios de justicia surgen de una deliberación en condiciones de imparcialidad, es decir, desde la “posición original” y bajo un “velo de ignorancia” acerca de las circunstancias que realmente van a acontecer a los participantes en la deliberación (Rawls, 1971).

Para Habermas (1987) [1981], los principios morales son fruto del consenso logrado mediante la deliberación racional en condiciones ideales de diálogo.

Estas teorías coinciden en que el criterio de lo moralmente correcto puede justificarse racional u objetivamente mediante consensos logrados en condiciones de libertad e imparcialidad de los individuos, dejando los sentimientos fuera de su procedimiento justificativo.

Desde estas corrientes, que enfatizan la racionalidad de las decisiones morales como seguimiento de reglas procedimentales, la subjetividad se convierte en un factor que resta imparcialidad. Esta línea de pensamiento conduce a la denominada *Machine Ethics* o *IA Ética*, dirigida a diseñar algoritmos que integren reglas de decisión moral (Anderson y Anderson, 2011). Bajo esta perspectiva, bastaría con descomponer el razonamiento ético en un conjunto de directrices objetivas para que la máquina emitiera juicios con contenido moral. Así, la *Machine Ethics* se ocupa de diseñar algoritmos que razonan de acuerdo con principios morales predeterminados y no conforme a criterios morales recurrentemente encontrados en los datos de entrenamiento. Desde la IA ética se propone incluir tales directrices o pautas éticas explícitas en los distintos modelos de IA, ya se usen en la medicina, la conducción autónoma o la publicidad (Moor, 2006). Por ejemplo, en un coche autónomo, se busca que el sistema sopesa valores como la protección de la vida humana y la prevención de daños materiales. Es más, se pretende que el coche autónomo pueda dilucidar, en el caso de

que un peatón se cruce inesperadamente en el trayecto del coche, la solución al conflicto que se plantea entre la seguridad del peatón y la seguridad de los pasajeros (Allen, Smit, & Wallach, 2005).

Las propuestas de la IA ética, atractivas a primera vista, chocan con el relativismo y con el pluralismo de valores señalado por Isaiah Berlin, según el cual existen distintos fines morales legítimos que pueden entrar en conflicto y no siempre es posible conciliarlos mediante un único criterio racional (Berlin, 1991) [1953]. Esta visión contrasta con la aspiración de diseñar sistemas de IA que adopten principios universales, pues la diversidad cultural y los intereses divergentes dificultan la implantación de reglas morales de validez general. A ello se suma el particularismo moral, que rechaza la existencia de pautas aplicables a cualquier contexto y defiende la atención a las circunstancias específicas de cada caso (Dancy, 1993). En consecuencia, o la IA Ética avanza hacia una mayor flexibilidad que incorpore esta complejidad y pluralidad de valores, o se enfrentará a enormes dificultades para materializar sus propuestas.

Otras aportaciones sobre la IA Ética se contienen *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way* (Dignum, 2019) y *The Oxford Handbook of Ethics of AI* (Dubber *et al.*, 2020), ofreciendo estrategias para integrar valores morales en el desarrollo de algoritmos. Asimismo, Nyholm (2020) examina el ajuste de diversas teorías éticas a situaciones concretas.

6.3. La coherencia como clave de la racionalidad práctica y la simulación de un discurso coherente por los LLMs

El concepto de *coherencia* desempeña un papel clave en la racionalidad práctica, exigiendo congruencia entre principios éticos, juicios morales particulares y las acciones concretas de los individuos. Esta exigencia se ha articulado principalmente desde dos enfoques diferenciados: por un lado, mediante teorías éticas centradas en la universalizabilidad y *consistencia lógica de los juicios morales*, y, por otro lado, a través del enfoque de la *coherencia narrativa*.

Dentro del primer enfoque, destaca R. M. Hare (1991) [1963], cuyo prescriptivismo ético sostiene que los juicios morales son esencialmente prescripciones universalizables que exigen coherencia lógica interna y aplicabilidad universal. Para Hare, actuar racionalmente implica juzgar y decidir moralmente de tal modo que se eviten contradicciones lógicas y prácticas, garantizando que cualquier principio ético pueda sostenerse consistentemente en situaciones similares (Hare, 1991).

Un planteamiento próximo es el de Rawls (1995)[1971] en virtud del equilibrio reflexivo: la racionalidad práctica requiere una coherencia dinámica y reflexiva entre las intuiciones morales inmediatas, los principios éticos generales y los juicios sobre casos concretos. Esta coherencia se alcanza mediante ajustes recíprocos continuos, permitiendo corregir y modificar tanto principios generales como juicios particulares, hasta conseguir un equilibrio estable y argumentativamente sólido (Rawls, 1995).

El segundo enfoque, referido a la coherencia narrativa, es elaborado principalmente por Alasdair MacIntyre y Paul Ricoeur. MacIntyre (1987) [1981] sostiene que la racionalidad práctica es inseparable de la capacidad humana para narrar la propia vida como un relato coherente, orientado hacia determinados fines o virtudes. Por su parte, Ricoeur (1996) [1990] enfatiza la importancia de la narración como estructura clave de la racionalidad práctica, pues implica la capacidad de articular la propia identidad personal en una narrativa coherente, permitiendo atribuir significado ético a los eventos vitales y decisiones tomadas.

Ronald Dworkin (1994) [1993] lleva la coherencia narrativa al campo de las decisiones públicas, como las que tienen lugar en el Derecho. Decidir racionalmente implica coherencia narrativa con los principios que están en la base del Derecho. Según Dworkin, la coherencia argumentativa surge del esfuerzo por integrar decisiones específicas en una narrativa vital en la que los principios éticos son continuamente valorados y ponderados según su relevancia y peso moral en circunstancias específicas.

Asimismo, Neil MacCormick (2005) [1978], destaca la relevancia de la coherencia para garantizar la estabilidad argumentativa en el ámbito práctico y jurídico, considerando que la racionalidad de las decisiones morales y jurídicas se funda en buena medida en la coherencia interna de los argumentos que las sustentan.

En la actualidad, los modelos de lenguaje generativos (LLMs) han mostrado capacidad para simular esta coherencia argumentativa en sus respuestas éticas. Aunque carecen de genuina deliberación moral, imitan con éxito patrones de razonamiento coherente basados en datos previos, reflejando así una especie de “coherentismo moral simulado”. En esta línea, Korsgaard (2009) [1996] sugiere que la racionalidad práctica puede entenderse en parte como la capacidad de ofrecer respuestas consistentes frente a circunstancias variadas, lo cual refuerza la similitud entre el razonamiento moral humano y la simulación realizada por la inteligencia artificial.

7. ALCANCE Y LÍMITES DEL “RAZONAMIENTO JURÍDICO ARTIFICIAL”

Una vez establecidas algunas premisas sobre el funcionamiento de la IA y su alcance y límites a la hora de emular facultades racionales teóricas y prácticas, el objetivo es examinar su alcance en el razonamiento jurídico. La hipótesis de este epígrafe es que la forma en que se concibe el Derecho condiciona la respuesta acerca de las posibilidades y límites de la automatización de procesos jurídicos en general y, más concretamente, del uso de la IA en la aplicación judicial de normas jurídicas.

7.1. *La concepción realista del Derecho y la IA jurídica con función predictiva*

Los realistas jurídicos, especialmente los norteamericanos como Holmes (2012) (1897) o Frank (1930), sostienen que el Derecho no es un conjunto de normas preestablecidas en las que los jueces fundan sus decisiones. Antes bien, el Derecho es aquello que los jueces deciden mediante sus sentencias y fallos. Para el realismo jurídico, el estudio del Derecho tiene por objeto la realidad, empíricamente aprehensible, de las decisiones judiciales, así como los motivos psicológicos, políticos y sociales que están detrás de tales decisiones (contexto de descubrimiento). No es, pues, el objeto del realismo jurídico, ni la interpretación de los enunciados normativos, ni las operaciones racionales que permiten resolver los casos particulares de acuerdo con normas generales, ni la elaboración de conceptos, ni la contribución a una comprensión sistemática del ordenamiento jurídico mediante construcciones teóricas y dogmáticas. En cambio, una de las tareas centrales de la ciencia jurídica realista es *predecir* cómo se comportarán los tribunales ante determinadas controversias. Esta visión realista del Derecho, que encuentra continuidad en autores como Brian Leiter (2007), Richard A. Posner (2008), Frederick Schauer (2009), y Brian Tamanaha (2017), profundiza en la relación entre la práctica judicial efectiva y los fundamentos empíricos, psicológicos y sociológicos que subyacen a la decisión judicial.

A estos efectos, la IA puede considerarse una herramienta de primer orden. Como venimos indicando, la función más característica de la IA es la predicción, analizando patrones (regularidades estadísticamente relevantes) en hechos acaecidos con anterioridad. En el ámbito del Derecho, la IA, analizando las resoluciones judiciales, puede predecir aspectos como fallos en casos similares o criterios interpretativos usados por diferentes tribunales en la decisión de casos (Aletas *et al.*, 2016).

En la actualidad, en efecto, estas aplicaciones ofrecen importantes ventajas, especialmente a los profesionales de la abogacía, que pueden aportar celeridad y calidad a su estrategia procesal. Con todo, la calidad de la predicción depende de la fiabilidad de los datos de entrenamiento. Además, cabe subrayar que la IA jurídica predictiva puede basarse en un modelo bastante simple o, por el contrario, requerir un modelo muy complejo; simple, si se limita a predecir los fallos de los tribunales; muy complejo, si recae también sobre las *rationes decidendi*. Los errores de predicción pueden derivarse tanto del carácter incompleto de los datos de entrenamiento como de la falta de identificación por la IA de las propiedades relevantes de los casos contemplados en las decisiones judiciales.

La indudable utilidad de la IA jurídica predictiva para los profesionales de la abogacía podría contribuir a una práctica del Derecho demasiado desligada del significado que en ella ha de tener el principio de legalidad o el ideal de Estado de Derecho. La IA jurídica, exclusivamente predictiva, banaliza el carácter normativo del Derecho y su función de guía la conducta de ciudadanos y poderes públicos; refuerza el papel creador del Derecho por parte los jueces y debilita la crítica a prácticas judiciales escasamente fundamentadas. Asimismo, el uso de la IA jurídica predictiva fomenta visiones del Derecho vinculadas exclusivamente al contexto de descubrimiento de la producción judicial de normas.

Al abordar la relación entre el realismo jurídico y la inteligencia artificial predictiva resulta esencial distinguir el enfoque de la Escuela Genovesa. Esta corriente comparte con el realismo norteamericano la noción de que los jueces desempeñan un papel activo en la creación del Derecho. Sin embargo, a diferencia del realismo norteamericano, que se centra en la predicción de decisiones judiciales y en el contexto de descubrimiento, la Escuela Genovesa pone énfasis en el análisis conceptual de los procesos interpretativos y aplicativos que evidencian dicha creación judicial (Guastini, 2017 [1996]; 2014).

En este marco, la IA, con sus capacidades avanzadas de procesamiento del lenguaje natural, podría ser una herramienta valiosa para identificar, describir y clasificar patrones en los procesos interpretativos y aplicativos del Derecho, facilitando una comprensión más profunda de cómo se manifiesta la creación judicial del Derecho en la práctica.

En síntesis, IA jurídica predictiva, tan usada en nuestros días, refuerza la concepción del Derecho propia del realismo jurídico. Desde este punto de vista, el razonamiento jurídico artificial se ciñe al análisis del resultado de las resoluciones judiciales, empobreciendo la dimensión normativa del Derecho.

7.2. La tesis positivista de la discrecionalidad judicial y el fracaso de los sistemas expertos

Las tesis formalistas del Derecho impregnaron el ámbito de la inteligencia artificial jurídica en la etapa del auge de los sistemas expertos. Como se señaló anteriormente, estos combinan una amplia base de conocimiento jurídico (legislación, precedentes y casos) en lenguaje formalizado, con un motor de inferencia lógica. Gracias a esta estructura, los sistemas expertos son capaces de resolver con fiabilidad problemas jurídicos, siempre que estos coincidan exactamente con los contenidos en la base de conocimiento, fracasando en los casos cuyas propiedades no se correspondían de manera exacta con ninguna formalización de norma o caso tipo.

Así, la IA de sistemas expertos resulta extremadamente limitada a la luz de cualquier una concepción del Derecho no formalista. Kelsen, en su segunda edición de la *Teoría Pura del Derecho* (1960), sostiene que las normas jurídicas no determinan por completo la decisión judicial, sino que configuran un marco con diversas soluciones posibles; así, el juez dispone de un margen de discrecionalidad acotado a las opciones permitidas por el ordenamiento.

Hart (1963)[1961] indica que los enunciados jurídicos poseen un “núcleo de certeza”, que recoge sus casos típicos, pero también una “zona de penumbra”, en la que se encuentran los casos fronterizos y es en ella donde el Derecho es *indeterminado*: no es posible afirmar con el valor de verdad o falsedad que una norma específica resuelve el caso, porque las proposiciones jurídicas aplicables carecen de tales valores en esa situación.

La indeterminación del lenguaje jurídico y algunas más son las causas, a juicio de Aguiló (2024), del fracaso de los sistemas expertos jurídicos: primera, en efecto, el problema del “no signo”: el Derecho se expresa en lenguaje natural y tiene un potencial interpretativo ineludible debido a la ambigüedad semántica y sintáctica, y a la vaguedad; segunda, de este potencial, los juristas no se quieren ver privados; tercera, la “no monotonía” o derrotabilidad del razonamiento jurídico: el razonamiento jurídico es “no monotónico”, es decir, añadir más información a las premisas iniciales, puede conllevar un cambio en las conclusiones; y cuarta, un contexto conflictivo, en lugar de cooperativo, para adoptar la decisión.

La cuestión es entonces si una IA más flexible, como la IA generativa de LLMs, puede tener utilidad en el razonamiento jurídico, comprendido de esta forma compleja.

7.3. *La concepción argumentativa del Derecho y las posibilidades de los LLMs*

La interpretación y aplicación de normas requiere generalmente un razonamiento que va más allá de la literalidad, incluso en casos aparentemente claros, y en ocasiones habrán de activarse las excepciones o *defeaters*. Ahora bien, esta inevitable *discrecionalidad* presente en la aplicación del Derecho no equivale a la *arbitrariedad*. La discrecionalidad es el margen de elección que el ordenamiento jurídico deja debido a la textura abierta del lenguaje o a la necesidad de hacer balance de razones. La arbitrariedad, en cambio, implica una decisión injustificada, caprichosa o desvinculada de las razones jurídicas. De ahí la importancia de la *concepción argumentativa del Derecho* y su combate frente a decisiones carentes de justificación. En un Estado de Derecho esa función creativa está limitada por la obligación de justificar las decisiones (Atienza, 2013).

La IA basada en el Big Data y el Machine Learning —indica Aguiló (2024)— ha transformado la forma de trabajar con datos en lenguaje natural, permitiendo superar los bloqueos de los sistemas expertos. Puede detectar patrones de interpretación y aplicación de la ley, a partir de un análisis masivo de sentencias judiciales.

Como se ha visto al analizar el funcionamiento de la IA que aprende (ML; DL), el algoritmo analiza ingentes volúmenes de datos para identificar regularidades o patrones estadísticamente relevantes (Brown *et al.*, 2020). Y con el procesamiento del lenguaje natural (PLN), los modelos de lenguaje a gran escala (LLMs) se entrenan con grandes conjuntos de textos para reconocer cómo se ordenan y combinan las palabras, pero también cómo se asocian los términos a significados en contextos determinados. Las herramientas detectan correlaciones matemáticas, pero muestran una aparente comprensión y capacidad de uso del lenguaje.

La IA jurídica se entrena con datos jurídicos en lenguaje natural procedentes de las fuentes del Derecho, que deben ser completas y fiables (bases de legislación y jurisprudencia de carácter oficial y actualizables), pero también cuentan en su entrenamiento con reglas de funcionamiento, como la jerarquía o la dimensión cronológica de las normas. En particular, el entrenamiento incluye “anotaciones” sobre el rango normativo (p. ej., una ley prevalece sobre un reglamento) y sobre la autoridad de las sentencias (el Tribunal Supremo crea criterios vinculantes para el resto de los tribunales, mientras que la doctrina de tribunales inferiores o la doctrina científica, puede servir de criterio interpretativo-aplicativo, pero no tiene el mismo valor).

Como se ha visto anteriormente, la IA puede ofrecer respuestas *predictivas*, es decir, la solución más probable a la vista del análisis de los resultados de las sentencias previas, pero también respuestas *normativas*, es decir, la solución más coherente con los patrones normativos aprendidos o incorporados al algoritmo; o con la base de conocimiento suministrada por el usuario; o con las instrucciones —legal promptig— recibidas. En palabras de Aguiló: la inteligencia artificial tiene que poder distinguir entre *patrones predictivos* y *reglas justificativas*. Los patrones predictivos son regularidades fácticas que la IA puede identificar en las decisiones judiciales anteriores, mientras que las reglas justificativas son normas que explican y legitiman esas decisiones. Para que la IA pueda ser eficaz en el ámbito judicial —señala el autor— debe ser capaz de diferenciar entre estos dos tipos de normas. Además, dado que los patrones (ya sea con función predictiva o con función normativa) se infieren principalmente de la jurisprudencia, es esencial que la cultura jurídica acepte los precedentes o normas de origen judicial (2024).

En la práctica, como destaca Solar Cayón, diversos sistemas de inteligencia artificial (IA) comienzan a emplearse, de forma incipiente pero cada vez más extendida, para asistir o incluso automatizar parcialmente la toma de decisiones judiciales. Cabe resaltar experiencias internacionales como las llamadas “cortes digitales” en China o el sistema argentino PROMETEA. En el ámbito europeo y español, se van introduciendo herramientas basadas en PLL para proponer resoluciones judiciales preliminares. Sin embargo, el Real Decreto-ley 6/2023 establece que cualquier “borrador” de resolución generado automáticamente debe pasar por el control y la validación del juez o tribunal. La IA no puede “decidir” un caso: no sustituye la función jurisdiccional ni la potestad de juzgar, que siguen correspondiendo a la autoridad competente. Además, el Real Decreto-ley subraya la importancia de establecer mecanismos de transparencia y explicabilidad que garanticen la independencia judicial, los derechos de las partes y las debidas garantías procesales (Solar Cayón, 2025, pp. 315-330).

En resumen, la IA que aprende lenguaje natural es mucho más adaptativa y, por tanto, mucho más útil en el razonamiento jurídico interpretativo-aplicativo que los antiguos sistemas expertos, que requerían un encaje perfecto entre el caso y la base de conocimiento. Sin embargo, su “margen de discreción” puede llegar a ser muy grande por lo que se requeriría que los LLMs jurídicos, siempre concebidos como solo como asistentes, progresaran en la explicación del razonamiento que les lleva a la solución dada.

Como señala Floridi (2024), la IA constituye una “agencia sin inteligencia”, ya que produce resultados propios del razonamiento humano sin poseer las características internas de la mente humana. Esta capacidad es prometedora para el razonamiento jurídico, que concibe el ordenamiento

como un sistema de reglas cuyo vehículo de transmisión es el lenguaje. No obstante, al igual que ocurre con los jueces, la IA legal, que no funciona de modo lógico, presenta un margen de discreción en los casos difíciles; discreción sesgada por los datos de entrenamiento y por la interacción del humano a través del *legal prompting*. Al igual que sucede con las decisiones de los operadores jurídicos, la IA realiza de forma bastante adecuada la justificación interna, pero quedan dudas sobre el modo en que la IA lleva a cabo la justificación de las premisas, que, aun cuando fueran explícitas gracias a los actuales modelos de razonamiento, dependerán de: a) los patrones de fundamentación que la IA haya encontrado en su entrenamiento; b) o de tales patrones unidos a la influencia de la interacción IA-máquina a través de *legal prompting*. Posiblemente, estos dos aspectos de la IA legal sean relevantes para la teoría de la argumentación jurídica.

7.4. *La teoría no positivista del Derecho como integridad y la imposibilidad de una IA jurídica*

Tal como expone Dworkin, una sentencia no se legitima solo por ajustarse a determinadas reglas formales, sino por articularse de forma coherente con los principios de la comunidad. Conocer el Derecho va más allá de manejar sus elementos formales o semánticos; exige participar en la práctica social en la que se enmarca y comprender los valores que legitiman el ordenamiento (Dworkin, 1986).

En el razonamiento jurídico intervienen reglas, pero también principios. En un caso particular, es posible que *prima facie* proceda la aplicación de una regla, pero esta puede ser derrotada si el principio moral en el que se fundamenta no resulta aplicable al caso, porque *all things considered*, debe triunfar el principio contrapuesto. Este razonamiento ponderativo destaca la dimensión moral consustancial al Derecho (Dworkin, 1986).

Desde esta perspectiva, la inteligencia artificial jurídica debería integrar no solo reglas jurídicas, sino también principios morales en su proceso decisorio. Este aspecto, como hemos visto en relación con la Machine Ethics, sí es posible, porque tales argumentos morales son detectables por los LLMs analizando el uso del lenguaje. Pero debería además ser capaz de comprenderlos, participando de la comunidad política e interiorizando sus fines colectivos. Este segundo aspecto no es posible. Un sistema de IA aplica normas sin comprender sus fundamentos morales, emitiendo decisiones “aparentemente coherentes con los principios morales” pero en realidad carecerían de dimensión moral.

Si el razonamiento jurídico demanda el conocimiento profundo de los valores morales y la inmersión en una forma de vida compartida, resulta difícil que una máquina, desprovista de conciencia, intencionalidad y experiencia subjetiva o emociones sea capaz de razonar jurídicamente. Aunque la IA disponga de una gran capacidad de análisis de datos y pueda ofrecer la solución más apropiada al caso y más coherente con patrones sobre principios morales fundamentales en la sociedad, le falta la experiencia y el conocimiento que adquiere el ser humano al sentir y vivir en sociedad.

A la luz de esta perspectiva, el razonamiento jurídico va mucho más allá de deducir consecuencias normativas de modo lógico, pero también de encontrar regularidades en la interpretación del Derecho o criterios morales adecuados para resolver un caso particular. El razonamiento jurídico está dirigido a legitimar la decisión ante la sociedad y salvaguardar la coherencia moral del ordenamiento, mediante una participación consciente e intencionada en la sociedad, lo que, salvo en un escenario de Inteligencia Artificial General, es imposible.

8. RECAPITULACIÓN

A lo largo del trabajo se ha tratado de llegar a algunas conclusiones sobre el alcance y los límites de la inteligencia artificial en el ámbito específico del razonamiento jurídico.

En primer lugar, se ha constatado que los grandes modelos de lenguaje (LLMs) son capaces de simular distintos tipos de razonamiento humano: deductivo, inductivo, abductivo y analógico. Sin embargo, esta simulación está condicionada por la naturaleza estadística e imitativa del aprendizaje automático, lo que implica una ausencia de comprensión genuina y de razonamiento causal y conceptual profundo.

En segundo lugar, en lo que se refiere al razonamiento práctico o moral, la IA puede ser exitosa si su criterio es objetivo y cuantificable, como sucede desde la perspectiva del utilitarismo o del Análisis Económico del Derecho. Más dudas arroja la IA inspirada en las teorías procedimentales de Rawls y Habermas. Finalmente, las teorías éticas basadas en emociones, empatía o experiencias subjetivas —como el sentimentalismo moral o la ética del cuidado— se sitúan fuera del alcance de la IA.

En tercer lugar, la aplicabilidad de la IA en el ámbito del razonamiento jurídico está estrechamente vinculada a la concepción del Derecho desde la cual se aborda. Así, la IA predictiva encuentra una correspondencia directa con el enfoque realista, centrado en las decisiones judiciales efectivamente adoptadas por los tribunales. Por otro lado, los antiguos sistemas expertos,

basados en reglas formales, encajan con el positivismo formalista, y encuentran importantes limitaciones debido a la textura abierta del lenguaje y la discrecionalidad inherente al razonamiento judicial.

Desde una concepción argumentativa del Derecho, los modelos IA generativa presentan un potencial significativo, ya que pueden detectar patrones interpretativo-aplicativos en la legislación y jurisprudencia, facilitando así la elaboración de argumentos jurídicos coherentes y justificados. Sin embargo, estos sistemas, al igual que los jueces, presentan una elevada discrecionalidad, resultando indispensable mejorar su transparencia y capacidad explicativa mediante técnicas específicas como el legal prompting.

En cuarto lugar, se ha destacado que desde la teoría del Derecho como integridad propuesta por Ronald Dworkin, la IA presenta serias limitaciones, dado que no puede involucrarse conscientemente en los principios morales fundamentales de una comunidad jurídica. Aunque pueda simular decisiones coherentes desde una perspectiva formal, carece de auténtica participación subjetiva en la vida moral y social, lo que le impide realizar razonamientos jurídicos en sentido genuino.

Finalmente, en consonancia con lo establecido por el Reglamento Europeo de Inteligencia Artificial (Reglamento UE 2024/1689), debe enfatizarse la importancia de la implicación institucional en el uso de la IA por la jurisdicción. Esta regulación reconoce expresamente el elevado riesgo que implica y exige mecanismos institucionales efectivos para verificar continuamente su fiabilidad técnica, argumentativa y ética.

La IA, por tanto, posee notables capacidades para auxiliar a la jurisdicción, hallando patrones argumentativos aprendidos a través de un proceso de entrenamiento, que debe estar garantizado desde el punto de vista de su confiabilidad. Con todo, lo explica muy bien Moral Soriano: la IA posee carácter instrumental y subordinado a la deliberación y justificación propias del razonamiento jurídico. En sus palabras: “La Ley de Hume establece un serio límite de la IA en el Derecho: es imposible derivar enunciados prescriptivos exclusivamente de premisas descriptivas; por tanto, es imposible derivar una decisión jurídica (juicio de deber) solamente de una razón o premisa descriptiva. Esto significa que los sistemas de ADM (toma de decisiones automatizadas) son conceptualmente incapaces de justificar / argumentar / fundamentar decisiones jurídicas” (Moral Soriano, 2023, p. 159).

Es decir, los límites de la IA se encuentran en su falta de comprensión genuina de los conceptos y relaciones causales que pueden estar presentes en la justificación externa e interna de las decisiones que adopte.

REFERENCIAS BIBLIOGRÁFICAS

- Aguiló, J. (2024). Notas Sobre Inteligencia Artificial y Decisión Judicial. *Revista Jurídica De Les Illes Balears*, 26, 107-26. doi:10.36151/RJIB.
- Aletras, N., Tsarapatsanis, D., Preo iuc-Pietro, D., y Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: A natural language processing perspective. *PeerJ Computer Science*, 2, <https://peerj.com/articles/cs-93/>
- Aliseda, A. (2006). *Abductive Reasoning: Logical Investigations into Discovery and Explanation*. Springer.
- Anderson, M., y Anderson, S. L. (2011). *Machine Ethics*. Cambridge University Press.
- Allen, C., Smit, I., y Wallach, W. (2005). Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology*, 7 (3), 149-155. <https://doi.org/10.1007/s10676-006-0004-4>
- Atienza, M. (2013). *Curso de argumentación jurídica*. Trotta.
- Bender, E. M., y Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. En *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185-5198). <https://doi.org/10.18653/v1/2020.acl-main.463>
- Berlin, I. (1991) [1953]. *El erizo y el zorro: Un ensayo sobre la visión de la historia de Tolstói*. Alianza.
- Bostrom, N. (2016). *Superinteligencia: Caminos, peligros, estrategias*. Teell Editorial.
- Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *ArXiv preprint arXiv:2005.14165*. Disponible en: <https://arxiv.org/abs/2005.14165>
- Calvo, J. (2024, 13 de agosto). MoA (Mixture of Agents). *Europeanvalley*. <https://europeanvalley.es/noticias/moa-mixture-of-agents/#:~:text=MoA%20se%20basa%20en%20la,s%C3%B3lida%20que%20cualquier%20experto%20individual>
- Chalmers, D. (2010)[1996]. *La mente consciente: En busca de una teoría fundamental* (Trad. de J. A. García). Siglo XXI.
- Collins, T., Dasgupta, I., & Eisape, D. (2022). Truthful AI: Assessing the factual accuracy of language models. *Journal of Artificial Intelligence Research*, 74, 1027-1064. <https://doi.org/10.1613/jair.1.13489>
- Cotino Hueso, L. (2024). ¿Qué es “inteligencia artificial” para el reglamento? Análisis, delimitación y aplicaciones prácticas. En L. Cotino Hueso & P. Simón Castellano (dirs.), *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea*, pp. 113-121.
- Dancy, J. (1993). *Moral Reasons*. Oxford: Blackwell.

- Dasgupta, I., Eisape, D., & Collins, T. (2023). Reasoning with Large Language Models. *AI & Society*, 38(4), 1234-1256. <https://doi.org/10.1007/s00146-022-01456-x>
- Dennett, D. (1991). *La conciencia explicada*. Barcelona: Paidós.
- Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Cham: Springer.
- Domingos, P. (2015). *The Master Algorithm*. Nueva York: Basic Books.
- Dreyfus, H. (1992). *Lo que las computadoras no pueden hacer*. Barcelona: Gedisa.
- Dubber, M., Pasquale, F., & Das, S. (eds.) (2020). *The Oxford Handbook of Ethics of AI*. Oxford: Oxford University Press.
- Dworkin, R. (1994) [1993]. *El dominio de la vida: Una discusión acerca del aborto, la eutanasia y la libertad individual* (Trad. Ricardo Caracciolo y Víctor Ferreres). Ariel.
- Dworkin, R. (1986). *El imperio de la justicia*. Barcelona: Gedisa.
- Eisape, D., Dasgupta, I., & Collins, T. (2023). Analytic reasoning in language models: possibilities and limitations. *Artificial Intelligence Review*, 56, 349-376. <https://doi.org/10.1007/s10462-022-10123-7>
- Floridi, L. (2024). *La cuarta revolución: Cómo la infósfera está reconfigurando la realidad humana*. Madrid: Taurus.
- Floridi, L., & Chiriatti, M. (2020). GPT-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30(4), 681-694. <https://doi.org/10.1007/s11023-020-09548-1>
- Freedman, D., Pisani, R., & Purves, R. (2007). *Statistics*. Nueva York: Norton.
- Gardner, H. (1987). *La nueva ciencia de la mente: historia de la revolución cognitiva*. Barcelona: Paidós.
- Gilligan, C. (1982). *In a Different Voice: Psychological Theory and Women's Development*. Cambridge: Harvard University Press.
- González Lagier, D. (2024). *Inteligencia artificial y razonamiento probatorio. Una conversación con ChatGPT-4o plus*. En prensa.
- Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. Cambridge: MIT Press.
- Guastini, R. (2017 [1996]). *La sintaxis del Derecho*. Madrid: Marcial Pons.
- Habermas, J. (1987 [1981]). *Teoría de la acción comunicativa*. Madrid: Taurus.
- Han, X., et al. (2024). Non-monotonic reasoning with language models. *Natural Language Engineering*, 30(2), 265-288. <https://doi.org/10.1017/S1351324923000192>
- Hare, R. M. (1991) [1963]. *Libertad y razón* (Trad. C. Mellizo). Alianza Editorial.
- Hart, H. L. A. (1963) [1961]. *El concepto de Derecho* (Trad. de G. Carrió). Abeledo-Perrot.
- Holmes, O. W. (2012) [1897]. *La senda del Derecho* (Trad. J. I. Solar Cayón). Marcial Pons Ediciones Jurídicas y Sociales.
- Huergo Lora, A. (ed.), & Díaz González, G. M. (coord.). (2024). *The EU Regulation on Artificial Intelligence: A commentary*. (Public Administration at the

- Boundaries. Studies and Perspectives on Evolving Public Law, n.º 9, Wolters Kluwer-CEDAM.
- Jackson, P. (1990). *Introduction to Expert Systems*. Addison-Wesley.
- Kelsen, H. (1982) [1960]. *Teoría pura del derecho* (Trad. de R. J. Vernengo, 2.ª ed.). Universidad Nacional Autónoma de México.
- Korsgaard, C. M. (2009) [1996]. *Las fuentes de la normatividad* (Trad. J. L. Pérez de la Fuente). Trotta.
- McCarthy, J., Minsky, M., Rochester, N., & Shannon, C. (1955). *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Disponible en: <http://www-formal.stanford.edu/jmc/history/dartmouth/dartmouth.html>
- MacCormick, N. (2005) [1978]. *Argumentación jurídica y teoría del Derecho* (Trad. J. Ferrer Beltrán). Marcial Pons.
- MacIntyre, A. (1987) [1981]. *Tras la virtud: Un estudio sobre la teoría moral* (Trad. Amelia Valcárcel). Crítica.
- Mill, J. S. (1984) [1863]. *El utilitarismo* (J. Rubio Carracedo, Trad.). Alianza Editorial.
- Moor, J. H. (2006). Nature, Importance, and Difficulty of Machine Ethics. *IEEE Intelligent Systems*, 21(4), 18-21. <https://doi.org/10.1109/MIS.2006.80>
- Moral Soriano, L. (2023). Criaturas empíricas en un mundo normativo: la inteligencia artificial y el derecho. *Revista de Derecho Público: Teoría y Método*, 7, 151-174. Marcial Pons Ediciones Jurídicas y Sociales. <https://www.revistas-marcialpons.es/revistaderechopublico/article/view/criaturas-empiricas-en-un-mundo-normativo>
- Nagel, T. (1974). What is it like to be a bat? *Philosophical Review*, 83, 435-450.
- Nagel, T. (2012). *Mente y cosmos*. Paidós.
- Penrose, R. (1989). *La nueva mente del emperador*. Grijalbo Mondadori.
- Popper, K. (1962) [1934]. *La lógica de la investigación científica* (Trad. V. Sánchez de Zavala). Tecnos.
- Posner, R. A. (2023) [1972]. *El análisis económico del Derecho*. Fondo de Cultura Económica.
- Putnam, H. (1967). Psychological predicates. En W. H. Capitan y D. D. Merrill (eds.), *Art, Mind, and Religion* (pp. 37-48). Pittsburgh: University of Pittsburgh Press.
- Qiu, Y., et al. (2024). Inductive reasoning capabilities of modern language models. *Journal of AI Research*, 75, 512-540.
- Rawls, J. (1971). *A Theory of Justice*. Cambridge, The Belknap Press of Harvard University Press.
- Rees, T. (2022). *New Forms of Intelligence*. MIT Press
- Rees, T. (2019). *After Ethnos*. Durham: Duke University Press.
- Ricoeur, P. (1996) [1990]. *Sí mismo como otro* (Trad. A. Neira). Siglo XXI.
- Russell, S. (2015). Unifying logic and probability. *Communications of the ACM*, 58(7), 88-97. <https://doi.org/10.1145/2699411>
- Searle, J. (1980). Minds, Brains, and Programs. *Behavioral and Brain Sciences*, 3, 417-457.

Solar Cayón, J. I. (2025). *Inteligencia artificial jurídica e imperio de la ley*. Tirant lo Blanch.

Webb, T., Holyoak, K. J., & Lu, H. (2023). Analogy in artificial intelligence. *Annual Review of Psychology*, 74, 543-569.

LEGISLACIÓN

Real Decreto-ley 6/2023, de 29 de diciembre, por el que se aprueban medidas urgentes para la ejecución del Plan de Recuperación, Transformación y Resiliencia en materia de servicio público de justicia (BOE núm. 312, de 30 de diciembre de 2023). <https://www.boe.es/eli/es/rdl/2023/12/19/6/con>

Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial y por el que se modifican los Reglamentos (CE) n. 300/2008, (UE) n. 167/2013, (UE) n. 168/2013, (UE) 2018/858, (UE) 2018/1139 y (UE) 2019/2144 y las Directivas 2014/90/UE, (UE) 2016/797 y (UE) 2020/1828 (Reglamento de Inteligencia Artificial) (DOUE núm. 1689, de 12 de julio de 2024) <https://www.boe.es/buscar/doc.php?id=DOUE-L-2024-81079>

Directrices de la Comisión relativas a la definición de sistema de inteligencia artificial establecida en el Reglamento (UE) 2024/1689 (Reglamento de IA). <https://digital-strategy.ec.europa.eu/en/library/commission-publishes-guidelines-ai-system-definition-facilitate-first-ai-acts-rules-application>.