https://doi.org/10.18800/iusetveritas.202502.011

IA generativa, razón práctica y Derecho: ¿Pueden los modelos generativos de lenguaje realizar verdaderos razonamientos prácticos?(*)(**)(***)

Generative AI, practical reason, and Law: Can generative language models perform genuine practical reasoning?

Alonso Ramiro Begazo Cáceres(****)
Universidad Católica San Pablo (Arequipa, Perú)

Resumen: Los modelos de inteligencia artificial, en especial los generativos de lenguaje, han irrumpido en la discusión académica con mucha intensidad. Estos modelos permiten la interacción con un sistema entrenado para la construcción algorítmica de respuestas dentro de una dinámica dialéctica, permitiéndonos absolver consultas, no solo de naturaleza conceptual, sino también de naturaleza resolutiva de problemas.

En el presente artículo, desde una perspectiva de la razonabilidad práctica de raíz aristotélica, argumentaremos por qué esta herramienta que se presenta como "inteligente" es incapaz de realizar verdaderos procesos de razonamiento práctico para orientar la acción por su ausencia de propiedades racionales, sobre todo en aquellos ámbitos directivos de la conducta humana vinculados a los bienes jurídicos que componen el Derecho.

Palabras clave: Inteligencia Artificial – Generativa - Grandes Modelos de Lenguaje - Razón Práctica - Razonamiento Jurídico – Reflexividad - Autoconsciencia – Autoimplicación – Limitaciones de la Inteligencia Artificial – Derecho Administrativo – Perú

Abstract: Artificial intelligence models, especially language-generating ones, have burst onto the academic scene with great intensity. These models allow interaction with a system trained to algorithmically construct responses within a dialectical dynamic, allowing us to answer questions not only of a conceptual nature but also of a problem-solving nature for various types of problems.

In this article, from the perspective of practical reasonableness with Aristotelian roots, we will argue why this tool, which presents itself as "intelligent," is incapable of carrying out true practical reasoning processes to guide action due to its lack of rational properties, especially in those areas governing human behavior linked to the legal rights that constitute law.

Keywords*:* Generative Artificial Intelligence – Large Language Models – Practical Reason – Legal Reasoning – Reflexivity – Self-awareness – Self-involvement – Limitations of Artificial Intelligence – Administrative Law – Perú

(*) Nota del Equipo Editorial: Este articulo fue recibido el 22 de octubre de 2025 y su publicación fue aprobada el 31 de diciembre de 2025.
(**) Una versión preliminar y más breve de este trabajo fue presentada en las "XVI Jornadas Internacionales de Derecho Natural - Inteligencia artificial, justicia y Derecho" realizadas en la Universidad San Ignacio de Loyola (Lima, Perú) del 2 al 4 de octubre de 2024
(***) Agradezco los aportes del profesor José Carlos Chávez-Fernández Postigo de la Universidad Católica San Pablo y del profesor Luciano Laise de la Universidad de Piura en las referidas jornadas. También agradezco al profesor Marcus R. P. Boeira de la Universidade Federal do Rio Grande do Sul y al profesor Dante Delgado Alata de la Universidad Católica San Pablo con quienes tuve el agrado de compartir e intercambiar algunas de las ideas que se reflejan en este trabajo.
(****) Bachiller por la Universidad Católica San Pablo (Arequipa, Perú). Master en Filosofía por la Universidad Católica San Antonio de Murcia, Máster en Dirección Estratégica de Recursos Humanos en la Universidad CEU San Pablo, España. Profesor de Derecho Natural y Teoría del Derecho de la Universidad Católica San Pablo. ORCID: https://orcid.org/0000-0002-0166-1041. Correo electrónico: arbegazo@ucsp.edu.pe.

1. Introducción

El rápido desarrollo de la inteligencia artificial [IA], en este último tiempo, ha generado una importante transformación en diferentes ámbitos relevantes del mundo humano. Uno de los principales desarrollos ha estado centrado en el vertiginoso crecimiento de la aparente capacidad lingüística de estas herramientas, generando patrones parecidos al comportamiento lingüístico humano, evidenciando una capacidad similar a la producción de un discurso tal como lo hacen las personas.

Un ejemplo emblemático de este tipo de inteligencia artificial generativa, con altísimas capacidades de producción textual son los Grandes Modelos de Lenguaje. Estos sistemas de inteligencia artificial son capaces de producir texto en lenguaje humano desde una base enorme de datos de entrenamiento textuales y la incorporación de variados parámetros, ostentando grandes redes neuronales de procesamiento.

Las capacidades productivas de los grandes modelos de lenguaje funcionan a partir de modelos neuronales predictivos. Es decir, la forma en cómo se produce el texto requiere de la predicción de las palabras basándose en representaciones de incrustaciones previas. Esto causa que mientras logremos incorporar más datos al entrenamiento, serán más eficientes para identificar posibles relaciones entre palabras, buscando conexiones de implicación, contradicciones o neutralidad.

Se sabe que estos modelos adolecen de problemas de trazabilidad, es decir, al incluir volúmenes impresionantes de datos, una compilación millonaria de parámetros y un autoaprendizaje algorítmico predictivo para la construcción de sus respuestas; es sumamente difícil conocer cuáles son las razones por las que el modelo llego a determinada respuesta y no a otra. El problema de la trazabilidad no solo es una circunstancia compleja para una explicación externa, sino que los verdaderos desafíos estriban en los esfuerzos por lograr una explicación interna ¿tiene acaso los modelos generativos la capacidad de tener una representación de sí mismo que le permita saber qué hace y porqué lo hace?

Esto se torna de mucha relevancia, debido a que estos modelos son capaces de producir textos con la finalidad de orientar y dar directrices que pueden ser incorporados por los usuarios de estas herramientas para justificar su acción personal. Esto trae un escenario problemático, porque el avance tecnológico ha permitido el desarrollo de una herramienta capaz de formular en un lenguaje humano, exhortaciones prácticas que pueden impactar en la dirección de la vida de las personas sin la capacidad de, verdaderamente entender, porque cierta directriz u otra deben ser preferibles en una circunstancia particular.

Este tipo de procesos de razonamiento práctico o moral, no solo se circunscriben al ámbito de la ética, sino que son característicos también en otras disciplinas como el derecho y la política. En estos ámbitos, es tipo de herramientas son también capaces de ofrecer directrices, pero la circunstancia se torna aún más dramática, debido a que el derecho y la política requieren de directrices prácticas que puedas ser claramente justificables. Es decir, se necesitan de razones que justifiquen las elecciones, más aún cuando de por medio están en juego los derechos, las libertades y las obligaciones más importantes de las personas en la vida social.

Por ello, en la presente investigación intentaremos analizar críticamente si los Grandes Modelos de Lenguajes son capaces de realizar auténticos procesos de razonamiento práctico, en especial en aquellos ámbitos que están vinculados a la deliberación en torno a bienes jurídicos como es distintivo en el Derecho. Conviene aclarar que la perspectiva desde la cual se plantea realizar este análisis es desde un enfoque filosófico jurídico, por lo que no pretendemos, en este espacio, un trabajo detallado o minucioso sobre los Grandes Modelos de Lenguaje propio de las investigaciones en Computación, sino recurrir a ciertos tópicos que nos permitan justificar nuestro punto.

El itinerario discursivo estribará en abordar cuatro puntos. En la primera parte, intentaremos explorar brevemente qué es la IA generativa y que son los Grandes Modelos de Lenguaje para sentar ciertos conceptos fundamentales y comprender su funcionamiento. Esto será nuestro primer insumo para responder a la cuestión central del trabajo.

En la segunda parte intentaremos esbozar una idea de razón práctica, de raigambre aristotélica, desde la cuál explicaremos las implicancias e importancia de estos procesos aprehensivos, reflexivos, deliberativos y electivos en la conducción de la conducta moral del hombre, evidenciando que desde esta perspectiva, los procesos de razonamiento práctico tiene una íntima vinculación con el Derecho, debido a que en este último la actividad prudencial es central para la cautela de bienes jurídicos en orden a la justicia de la vida social. Esto nos permitirá comprender mejor los presupuestos de nuestro análisis.

En el tercer punto haremos una rápida revisión del algunos estudios e investigaciones que han tratado de medir la capacidad moral de los modelos generativos con la finalidad de recoger algunos ángulos de valor e inconsistencias de estas investigaciones de corte más fenoménico, con el propósito de evidenciar algunas de las limitaciones actuales de estos modelos. Con ello agregaremos un insumo final para la construcción de una respuesta a nuestra pregunta central.

Por último, ofreceremos una justificación detallada de por qué los grandes modelos de lenguaje son incapaces de realizar verdaderos razonamientos prácticos desde la perspectiva asumida. El centro de nuestra argumentación gravitará en el hecho de que propiedades esenciales del agente moral para el razonamiento práctico como la consciencia, la reflexividad, la implicación, la responsabilidad y la capacidad de justificación causan de que en ninguna circunstancia los modelos generativos puedan verdaderamente realizar razonamientos prácticos en general, ni tampoco razonamientos específicamente jurídicos; sino a lo sumo reproducir formulas, preceptos y directrices morales desprendidas de sus datos de entrenamiento.

Este análisis no se limita únicamente a clarificar las capacidades y limitaciones de los modelos generativos en la actividad directiva de la conducta humana en sus ámbitos morales y jurídicos, sino que también persigue desarrollar una reflexión profunda sobre lo que implica verdaderamente ser humano y las implicancias morales y sociales de claudicar ante la cesión de nuestra racionalidad a la "deliberación" algorítmica de manera irresponsable.

2. ¿Qué son los modelos generativos de lenguaje en el ámbito de la Inteligencia Artificial y cómo funcionan?

Antes de adentrarnos en abordar la capacidad de razonamiento práctico de los grandes modelos de lenguaje, nos parece pertinente sentar algunos conceptos clave para comprender como operan las IA Generativas; y con ello, exponer algunos tópicos básicos sobre el funcionamiento de los grandes modelos de lenguaje.

Atkinson señala que, en los primeros momentos de la investigación sobre la *IA, el enfoque se centraba principalmente en el desarrollo de sistemas basados en reglas predefinidos, que permitían desprender razonamientos y toma de decisiones, con una minuciosa supervisión de expertos en la configuración de estas reglas; manteniendo su espectro de acción en las reglas explícitamente programadas.* Pero acontece a nuestro tiempo la incursión de un nuevo paradigma ocasionado por lo que podemos llamar IA generativa, donde se erradica la necesidad de datos etiquetados. Estos sistemas de IA generativa logran formular razonamientos y lograr cierto tipo de toma de decisiones "aprendiendo" de forma independiente a través de una gestión algorítmica de enormes conjuntos de datos, que le permiten generar conclusiones y relaciones inherentes dentro de los datos. *(2025, p. 1).*

Este nuevo paradigma generativo permite que estos modelos de aprendizaje automático profundo logren la producción de nuevo contenido basado en la formulación propuesta por algún usuario, que se pueden realizar a través de descripciones o consultas elaboradas en estructuras propias del lenguaje natural. Las respuestas que puede ofrecer esta capacidad generativa permiten no solo la generación de textos escritos, sino también la producción de imágenes, vídeos, audio, música e incluso código informático (Alto, 2023).

Lo que hace posible la generación de este contenido novedoso es el enorme conjunto de datos de entrenamiento al que son sometidos estos sistemas, que están compuestos por formatos similares a los que pueden producir, es decir, texto, imagen, audios y otros; lo que permite configurar un sistema neuronal de aprendizaje profundo conocido como Redes Generativas Antagónicas. A través de esta red se filtran los datos recurriendo a un sistema que recompensa los aciertos y penaliza los errores para que puedan reconocerse y comprender mejor las intrincadas relaciones entre los datos, mediante un sistema de supervisión humana (Atkinson, 2025, p. 2).

La dinámica de estas redes antagónicas se compone, propiamente por dos redes: una generadora que crea nuevos datos y una discriminadora que evalúa su autenticidad; donde ambas redes buscan interactuar dualmente, con el propósito de lograr un contenido prácticamente indistinguible de los datos auténticos. (Atkinson, 2025, p. 2)

Este hecho encarna, quizá, uno de los desafíos más complejos que nos proponen estas nuevas tecnologías, y es que los productos que son capaces de generar confunden fácilmente a los usuarios como si fueran casi humanos. Es decir, producen texto, imágenes o audios que son casi indistinguibles a los que suelen ser el resultado de la actividad humana. Sentando

algunas notas esenciales sobre los alcances de la IA generativa, corresponde ir perfilando nuestra exploración a los grandes modelos de lenguaje.

Al respecto, Urdan y Marson señalan que el procesamiento del lenguaje natural es una de las ramas más importantes de la inteligencia artificial en el amplio mundo de la informática; y que buscan que estas tecnologías puedan desarrollar la capacidad de generar lenguaje similar al utilizado por las personas en la vida cotidiana, incluyendo la posibilidad de analizar los sentimientos, el reconocimiento de voz, la traducción automática, generación y resumen de texto, entre otros (2024; p.4).

Dentro del ámbito de los sistemas de procesamiento del lenguaje natural destaca la irrupción de los Grandes Modelos Lingüísticos [Large Language Model - LLM]; aunque cabe destacar que las raíces de los modelos de lenguaje generativo ya tienen cierto recorrido, encontrando exploraciones tempranas orientadas a crear sistemas de diálogo interactivos basados en el aprendizaje evolutivo a partir de datos web. (Ferreira y Atkinson, 2005) En la actualidad contamos con varios ejemplos notables de grandes modelos de lenguaje que incluyen GPT-3, GPT-4 y LaMDA. Estos LLM permiten la producción de un texto nuevo a través de procesos de muestreo estadístico, aprovechando los amplios conjuntos de datos de entrenamiento que sustentan su creación (Atkinson, 2025, p. 5).

Estos modelos se caracterizan por ser sistemas que se entrenan con una amplia cantidad de datos que incluyen textos de diferente índole, con la finalidad de que estos puedan ser utilizados para producir respuestas probables a través de una secuencia de palabras, permitiendo la generación de contenidos de características muy similares a lenguaje escrito humano.

Esto implica que las capacidades productivas de grandes modelos de lenguaje se basan en la presencia de modelos neuronales predictivos. Es decir, la forma en cómo se produce el texto requiere de la predicción de la siguiente palabra basándose en representaciones de incrustaciones previas a través de una técnica denominada Reconocimiento de la Implicación Textual, que mejora la comprensión de las relaciones entre palabras. Esto casusa que, a medida que se incorporan más datos al entrenamiento, el sistema analiza continuamente las relaciones entre palabras, buscando conexiones de implicación, contradicciones o neutralidad (Atkinson, 2025, P.6).

Por esta razón, la generación de textos en estos modelos requiere un trabajo predictivo con respecto de las palabras que tiene mayor probabilidad de aparecer después de las anteriores, considerando la base de información disponible de datos de entrenamiento. Es aquí donde surgen una importante limitación de estos modelos, que radica en el hecho de que es posible que los datos de entrenamiento contengan fuentes como Wikipedia materiales de ficción o teorías de conspiración que pueden terminar generando un texto plagado de información falsa o alucinaciones de diversa índole (Atkinson, 2025, P.7).

Por ello, en atención a esta deficiencia ha sido indispensable acompañar el desarrollo de estos modelos con una activa participación humana que permita mejores procesos de entrenamiento y de retroalimentación, que ha implicado que los entrenadores humanos interactúen con los modelos como usuarios y como asistentes de entrenamiento. Después de revisar las respuestas obtenidas como finalizaciones, se brinda retroalimentación a través de un sistema de recompensa. Esto es lo que se conoce como aprendizaje por refuerzo (Atkinson, 2025, P.7).

Esto permite que, con el trascurrir del tiempo, este sistema de recompensas perfeccione las respuestas en resultados más realistas; pero aun así no estamos libres de poder recibir contenido sesgado y que es imposible de ser verificado, como sería el caso de una apropiación indebida y no referenciada de ideas de un autor, lo que exige la necesidad de verificación rigurosa por parte de los usuarios la decidir usar estos sistemas dado el evidente problema de trazabilidad.

Esto evidencia que otro aspecto crucial del debate contemporáneo sobre esta tecnología implica la reflexión sobre el sentido de responsabilidad, dada la evidencia sobre la ausencia de responsabilidad directa de los contenidos producidos por estos modelos y la dificultad de trazabilidad de la información, debido al importante volumen de datos que genera y que debido a su complejidad algorítmica, los mismos autores de estas herramientas no conocen bien porque en determinadas circunstancias se puede generar una respuesta y no otra.

A modo de ilustrar esta exploración, dentro de estos grandes modelos de lenguaje, podemos identificar uno que se ha posicionado como representativo en el último tiempo, que no es otro que el conocido ChatGPT. Este sistema, desarrollado por la empresa *OpenIA*, se basa en un proceso de respuesta estadística que calcula la

elaboración de una respuesta siguiendo un mecanismo de probabilidad para la formulación de una secuencia de palabras inteligible. De este modo, el sistema presenta un texto como respuesta ante una consulta formulada por un usuario humano, donde la especificidad y claridad de la respuesta depende de la adecuación de la indicación realizada por el usuario (Urdan y Marson, 2024, p. 4).

El nombre de ChatGPT [*Chat Generative Pre-Trained Transformer*] se debe a su finalidad de ser un espacio de conversación con un sistema, caracterizado por ser generativo desde una base previamente entrenada. En ese sentido, este sistema logra la elaboración de sus respuestas a partir de un entrenamiento donde se incorpora información para poder acceder a palabras del idioma, con la finalidad de incrustar marcadores que permiten fijar vínculos para formar conjuntos de palabras, y progresivamente, lograr la construcción de textos complejos que pueden terminar siendo leídos de forma coherente (Navarro-Dolmestch y Fuentes-Loureiro, 2023, p. 3).

Tanto el modelo de GPT-3 y GPT-4 son grandes modelos de lenguaje capaces de procesar entradas de textos y permitir salidas basadas en texto. Como es característico en los grandes modelos de lenguaje, poseen una dinámica generativa marcada por una predicción algorítmica que permite la concatenación de tokens o palabras para la construcción de discursos ciertamente coherentes. Las principales diferencias entre estos dos modelos implican la amplitud de parámetros, los datos de entrenamiento y capacidad de razonamiento.

Atkinson advierte que una de las principales preocupaciones sobre la aplicación de los grandes modelos de lenguaje es la cuestión de la alineación; es decir, si las salidas que producen se ajustan a los valores humanos en términos de seguridad. (2025, p. 91) Esto evidencia uno de los principales riesgos de esta tecnología debido a las limitaciones del escenario del entrenamiento. Esto ha dado paso a que esta tecnología presente la generación de ciertas alucinaciones o sesgos en sus procesos, pudiendo causar que en el intento de construcción de una respuesta recurra a una información irreal o incorrecta, o peor aún, ofrezca instrucciones dañinas, que pueden atentar contra el bienestar humano, trasgrediendo ciertos principios morales o jurídicos en el manejo de la información.

Por ello, se han contemplado incluir en el diseño elementos que tiendan a reducir esos sesgos o alucinaciones a través del resguardo de tres cualidades: usabilidad, veracidad e inocuidad. Cuando hablamos de su usabilidad, nos referimos a su capacidad para seguir instrucciones y realizar tareas; cuando hablamos de veracidad nos referimos a su capacidad de proporcionar información objetiva y reconocer sus propias incertidumbres y limitaciones; y, por último, su inocuidad implica la capacidad de evitar respuestas tóxicas y negarse en participar de actividades peligrosas (Atkinson, 2025, P.7).

Estos parámetros arrojan ciertas respuestas eficientes cuando nos movemos en la generación de respuestas sobre un contenido teórico o especulativo, donde el tipo de información que ofrece es particularmente conceptual e informativa; pero cuando entra a los derroteros de construir respuestas vinculadas al razonamiento práctico y pretende esbozar respuestas orientadas a determinar que se debe hacer en una particular situación, comienza un escenario desafiante, debido a que el razonamiento práctico exige moverse en el mundo de lo contingente donde será necesario muchos más elementos que la concatenación de palabras a través de algoritmos.

Este caso particular, de respuestas potencialmente sesgadas, es relevante para nuestro análisis debido a que una de las tareas para las que se ofrece esta tecnología estriba en la resolución de problemas, que pueden ser de naturaleza moral o técnica. Esto implica, que esta herramienta podría ofrecer información y directrices a sus usuarios sobre ciertos ámbitos del conocimiento que podríamos definir como prácticos, como pueden ser el derecho y la política.

Desde esta perspectiva, los modelos generativos pueden crear diálogos y respuestas que pretendan ofrecer conclusiones de naturaleza directiva en el comportamiento humano, estableciendo preceptos que orientan prácticamente a usuarios a realizar determinadas conductas. En ese sentido, el uso de esta herramienta estaría suplantando una facultad distintiva para la ejecución de las tareas humanas, causando la degeneración cognitiva de las personas.

Pero además de estos riesgos, que en sí mismos son verdaderamente graves, es de nuestro interés centrarnos en analizar si estos modelos o herramientas verdaderamente pueden generar un proceso de razonamiento practico o moral en sentido estricto; o si solo recurren a tomar ciertos preceptos o mandatos que han modelado las fronteras de su entrenamiento. Porque si bien es cierto, nada debería suplantar la capacidad de razonamiento moral de las personas, menos aún lo debería realizar un sistema que, propiamente, procede de un modo infra racional, causando un escenario de mayor afectación a la aproximación moral de las personas.

3. ¿Qué es la razón práctica y cómo se vincula con el Derecho?

Antes de responder al cuestionamiento sobre la capacidad de razonamiento práctico de la herramienta analizada en cuestión, conviene esclarecer que entendemos por razón práctica. Conviene aclarar que el punto de partida desde el que abordaremos este desarrollo implicará suscribir una visión de la razón práctica, que Alexy ha señalado como "aristotélica", para diferenciarla de algunas aproximaciones contemporáneas con diferentes interlocutores, como las hobbesianas o kantianas (1996, p. 61); y que, a su vez, Massini ha denominado como iusnaturalista realista (2005, p.10).

García Huidobro indica que desde mediados del siglo pasado ha surgido un movimiento importante de "rehabilitación de la filosofía práctica", que ha implicado un resurgimiento de la clásica aproximación de razonabilidad práctica, aunque al mismo tiempo han sobrevenido otras aproximaciones distintas como las anteriores señaladas. (1993, p.22) Este movimiento encuentra su motivación en propiciar una crítica contra la pretensión de explicar al hombre y la sociedad sobre la base de la llamada "razón instrumental" y al empirismo positivista.

Desde este punto de vista, podemos afirmar que la razonabilidad práctica puede entenderse como el uso de la razón en la dirección de la *praxis* humana. Según Massini, esta aproximación de la dirección racional de la conducta implica en un primer sentido, la captación de los bienes humanos que por su naturaleza tienen la condición de fines del obrar; descubriendo el valor y su pertinencia para la ordenación humana, como también implica la deliberación y el establecimiento de los medios pertinentes para el logro de estos bienes humanos (Massini, 2005, p. 14).

Esta visión implica adoptar una idea antropológica que afirma que el ser humano es un ser inteligente. Es decir, es un ser capacidad para "leer en lo profundo" de la realidad pudiendo captar e interactuar con esas dimensiones trascendentales de la realidad como son la verdad, la bondad, la belleza y el ser. Esta capacidad de la inteligencia permite conocer la realidad y sus diversos aspectos que no tienen siempre las mismas cualidades.

Al existir cualidades diferentes en la realidad, la inteligencia capaz de la realidad, es capaz de explorar sus múltiples aspectos, ordenándolos en diferentes tipos de conocimiento. Estos diversos órdenes del conocimiento se corresponden a las diversas formas en cómo puede funcionar la inteligencia humana, y en ese sentido, descubrimos que existen realidades especulativas, prácticas y lógicas que pueden ser captadas por nuestra inteligencia. La razón humana, unas veces se dirige a la acción y otras hacia la contemplación de la verdad; no generando con ellos la existencia de pluralidad de inteligencias, sino la evidencia de que nuestra inteligencia puede funcionar de diversos modos (Selles, 2006, p. 297).

Considerando lo indicado, podemos señalar que el conocimiento que se produce del ejercicio de la razón práctica es lo que se denomina como conocimiento práctico. Cuando aludimos al conocimiento práctico nos referimos a aquel que versa sobre las acciones humanas, donde recurrimos a la "aplicación del conocimiento intelectual a la concreta situación en las que cada uno se encuentra". (Yepes, 1996, p. 104) Es decir que, su propósito no estriba en conocer y reflejar la realidad; no busca responder a la pregunta ¿qué es esto?, no busca quedarse en la mera teoría (Llano, 2015, p. 54), por el contrario, pretende establecer los cursos de la acción humana, por lo que no es separable de las acciones humanas donde encuentran su sentido, procurando responder a la pregunta ¿qué debo hacer? ante una determinada situación. Por ello podemos concluir que la razón teórica versa principalmente sobre cosas necesarias y la razón práctica se ocupa de cosas contingentes.

Es por ello que coincidimos con Taylor en que, desde una perspectiva amplia, la tarea del intelecto práctico es la iniciación del cambio que se terminará de operativizar por la voluntad, donde su inicio es identificable la captación de la verdad de acuerdo con el deseo recto; para posteriormente a través de la deliberación, valorar la conveniencia, la posibilidad y los medios pertinentes que permitan finalmente la elección, es decir, determinar cuál es el curso de la acción que deberá impulsar la voluntad (Taylor, 2014).

A partir de lo visto, podemos comprender que el conocimiento propio del orden moral no es un conocimiento ordenando que versa sobre realidades fijas, sino por el contrario lo categorizará como una orden que aún no es realidad; y que solo será constituirá como tal en el proceso en que el intelecto nutra de razones a voluntad para la acción, que conformarán la relación entre intención y elección.

Al señalar la existencia de funciones del intelecto se pone de manifiesto la diferencia que existe entre estas sus dos dimensiones; pero esto no nos debe llevar a concluir una desvinculación absoluta de la razón práctica

sobre la razón teórica, como tampoco ni una dependencia férrea de la razón práctica sobre los alcances de la razón teórica. El justo medio que consideramos conveniente implica entenderlo como una interdependencia, en donde lo teórico y lo práctico puede conocerse de modo independiente a causa de poseer principios independientes, pero a su vez vinculados, a razón de que el "fin" que es algo que se capta teoréticamente, pero al mismo tiempo, como señala Tomas de Aquino siguiendo a Aristóteles, se constituye en el primer principio del obrar.

La razón práctica permite la obtención de un conocimiento que está marcado por la intención de comprender y descubrir las razones que están detrás del comportamiento, así como sus condiciones. Este conocimiento implica, en gran medida, una indagación sobre la naturaleza, la dignidad y el destino de los hombres desde la perspectiva de los bienes humanos que son adecuados a las inclinaciones de la plenitud humana. (George, 2009, pp. 144-145) Desde ese sentido, el bien se constituye, de forma única, como aquello que permite mover la voluntad del hombre de forma racional hacia un fin, configurándose como una razón para la acción, comprometiendo al agente desde un punto de vista interno, causando que esa "verdad práctica" se implique en su comprensión de lo bueno y las posibles formas de alcanzarlo (Pereira, 2008, p. 5).

Esta verdad práctica, a diferencia de la verdad propia del mundo teórico que ofrece una verdad necesaria, es una verdad contingente que se asemeja a la idea verosimilitud o razonabilidad, que permite anticipar, mediante el intelecto, la posibilidad de que algo pueda conformarse a la razón y consecuentemente, a la realización de la acción y a su finalidad, es decir, la plenitud humana integral. Aunque cabe precisar, como lo hace Castaño, que:

> No se trata aquí de que sea solo la razón la que decide y construye la verdad o justicia de un acto, como ocurre en las propuestas constructivistas, sino de que la razón, constitutivamente vinculada a la realidad objetiva, formula a partir de los datos de esa misma realidad las directivas éticas de la vida humana (2013, p. 81).

En ese sentido, Taylor afirma que, desde una perspectiva de raíz aristotélica, la formulación de esta verdad práctica implicaría principalmente dos cuestiones: La validación de que lo que se cree sobre el bien sea verdadero y segundo, se necesita estar motivado para actuar de acuerdo con esas creencias. (2014) En ese sentido, el conocimiento moral requiere poder justificarse por parte del agente y al mismo tiempo se debe tener en consideración que su formulación no supone una deducción exógena sino más bien requiere de la "auto implicación" con esa formulación. Esta justificación e implicación necesaria del conocimiento práctico solo es posible en el escenario de ejecución de juicios razonados que implican la reflexión sobre los alcances, contenidos y consecuencias. Donde los atributos de reflexividad de la inteligencia resultan ser cruciales.

En razón a lo señalado, podemos caer en cuenta de que la capacidad operativa del hombre es inmensa, volviéndose permanentemente en un hacedor de sí mismo, realizándose a cada instante su propia vida. Cabe indicar brevemente, para evitar confusiones, que para los propósitos de este trabajo queremos centrarnos únicamente en la idea de práctico en el sentido estrictamente moral. Esta aclaración es porque cuando hablamos del mundo de la acción humana puede enfocarse desde el orden a moral, como también puede abordarse desde el orden de la producción, que distingue la obtención de un conocimiento que permite un "saber obrar" y por otro lado, la obtención de un "saber hacer" (Garcia, 2010, p. 81).

Está diferencia entre los ámbitos de la acción humana también suelen categorizarse desde las acepciones clásicas de lo factible y lo agible. El *faceré*, el "hacer", es una actividad inteligente que se ejerce sobre una materia perteneciente al mundo exterior: cortar, pintar, etc. En cambio, el *agere*, el "obrar", es la actividad que se ejerce dentro del hombre mismo: querer, odiar, etc.

A partir de lo señalado, podemos afirmar que la acción humana, en especial para nuestros fines la acción moral, termina siendo la respuesta a la atracción intelectual de un bien, en donde su consecución implica un primer movimiento de la razón. En ese sentido, la razón práctica termina siendo el primer elemento en la determinación de una acción, generando una representación ante uno mismo del curso de su acción, que permite que la propia decisión sea una decisión racional, antes que un impulso.

En ese sentido, las razones que una persona puede conocer para elegir y actuar son los bienes inteligibles, y que al ser varios los bienes, permitirán que los beneficios de cada acción puedan ser razonablemente concretados de diferentes maneras y, por lo tanto, ser elementos de realización de individuos y de asociaciones de forma diversa (Finnis, 2017, p.4).

Por esa razón, el afirmar la posibilidad del conocimiento del bien, nos abre a descubrir una noción de realización humana que es heterogénea, producto de las muchas formas

irreductibles de bienestar; que para nada busca negar que la naturaleza humana posea aspectos esenciales, sino que busca resaltar el particular carácter de complejidad de nuestra naturaleza, y, por ende, también de la actividad prudencial humana.

Como hemos visto hasta aquí, la razón práctica "no es un dispositivo de espejear verdades, sino de realizarlas. Es inventiva, proyectista. Fragua normas de conducta, traza formas de gobierno, arbitra medios para conseguirlas, planea instituciones, hilvana sistemas" (Gonzáles, 2006, p. 263). En ese sentido podríamos prever que, por su objeto, la razón práctica tiene una íntima vinculación con el derecho.

De hecho, cabe recordar que uno de los elementos más característicos de esta tradición es la existencia de una conexión fundamental entre la moral que la razón práctica humana puede reconocer y el derecho establecido por la autoridad que se traduce en la formulación de diversas prescripciones objetivas de justicia en la vida social (Chávez-Fernández, 2023, p. 41).

En ese sentido, cuando nos referimos a la acción moral humana, resulta evidente que esta no se circunscribe únicamente al espectro de la esfera individual y los compromisos con uno mismo, sino en muchos casos trasciende lo individual, y nos coloca de cara a la búsqueda de bienes que involucran a otros. Chávez-Fernández señala que, desde la perspectiva aristotélica, se insiste en que la racionalidad práctica es en gran medida fruto del ejercicio de la prudencia, en constante contacto, a través de la experiencia, con la contingencia de las realidades sociales en las que el razonamiento, práctico, en general; y jurídico, en particular, se concreta (2019, p. 149-150).

Esto puede ser palpable en la vida familiar y política, donde la persona se ve en permanente encrucijadas que requieren la intervención prudencial de la inteligencia y la formulación de cuestionamientos prácticos buscando identificar los bienes que se deben perseguir para encausar la conducta con respecto de los otros.

De hecho, la búsqueda de bienes en el contexto social va suponer una exploración compleja debió a la presencia del elemento de la libertad en un contexto social. A medida que es más complejo el contexto social, más complicada resulta la actividad deliberativa. Es por ello por lo que, en el ámbito de la comunidad política, la deliberación, que es esencial, no resulta nada fácil porque exige una particular involucración de todas las comunidades, individuales e institucionales, para fomentar un verdadero esfuerzo por sostener un dialogo que facilite la deliberación entre los fines y medios que se persigan como buenos para todos, es decir, los fines y medios para la obtención del bien común y la justicia.

Por ello, se puede identificar que, dentro del amplio espectro de los bienes morales sociales, existen algunos bienes básicos para la vida social y que pueden ser afectados por las esferas de dominio de los demás, que son los que normalmente señalamos como bienes jurídicos. Estos bienes jurídicos son procurados por la misma estructura intelectual que los bienes morales, porque además de catalogárseles como jurídicos son primeramente morales, es decir, adecuados al perfeccionamiento humano. Por ende, su identificación y procura va estar perseguida por la prudencia, por lo que pueden terminar constituyéndose como verdades prácticas, dentro de una comunidad humana.

Tanto como el legislador al comprometerse en la actividad de diseñar estructuras políticas y las normas jurídicas ejerce todas las capacidades de discurso racional y deliberación práctica (Duke, 2023, p. 98), del mismo modo los diversos operadores jurídicos al momento de interpretar su contenido y alcances para la resolución de los diferentes problemas de justicia que se generan en la vida social intentan comportarse como un agente prácticamente razonable con conocimiento del bien humano.

El bien jurídico, que dentro de la tradición aristotélica se ha definido como "lo justo" en el sentido de ser objeto de la justicia como acción o virtud, requiere de una intervención práctica de la inteligencia para su determinación, debido a que son muchas y variadas las relaciones de justicia que surgen de la vida en común. Por ello, como bien señala Gonzáles, es el hombre, quien, a través de su razón práctica, el que está en condiciones de reconocerlo, tanto en las relaciones entre particulares como en las relaciones entre el individuo y el Estado (2006, p. 273).

Si bien es cierto existen algunos bienes corresponden todo hombre por el hecho de poseer una naturaleza como la humana, existen algunos bienes, y medios para estos bienes, que pertenecen al mundo de lo "indiferente" o "contingente" de forma más radical, donde es indispensable que la razón práctica del legislador, los operadores jurídicos e inclusive los ciudadanos deban determinar su carácter práctico dentro de un contexto jurídico y sistémico concreto.

En ese sentido, queda claro que la actividad prudencial del jurista se enmarca dentro de un sistema con ciertas concreciones técnicas que no lo sitúan dentro del ejercicio de la razonabilidad práctica general, sino

dentro de un ejercicio de la razón práctica corregido por un variado conjunto de consideraciones técnicas e institucionales que rectifican el ejercicio prudencial dando paso a un especial caso de ejercicio de la razón práctica. Es decir, que en su actividad prudencial lo que persigue es determinar las razones de porqué determinada acción es buena, o más propiamente justa, considerando adicionalmente las precisiones o alcances del ordenamiento jurídico que enmarca su razonamiento.

Por ello podemos señalar que conocer el derecho requiere de un conocimiento práctico, (Hervada, 2011, p.17) y no porque sea irrelevante el conocimiento de ciertos aspectos técnicos o teóricos para garantizar su eficacia, sino porque principalmente conocerlo implica descubrir el bien o los bienes, y sus respectivos medios para garantizar su realización orientada a la plenitud personal en el contexto social.

Por esa razón es plausible afirmar que la pregunta por el derecho es una pregunta por una verdad práctica que admite múltiples y variadas formas de procurarse y requiere realizarse. Esto no implica que el conocimiento teórico y el conocimiento técnico no sea relevante; seguramente lo son, y mucho, pero de forma subordinada al descubrimiento del bien vinculado a la justicia. Y es que, si bien es cierto, la verdad en el derecho es una verdad moral; está cualidad es necesaria pero no suficiente, debido a que para su realización se necesitan de ciertas condiciones que permitan claridad, eficacia e idónea oportunidad para su procura.

Esto se traduce en la evidencia fenoménica que el mundo humano se caracteriza por considerar el papel activo de la razón práctica en el desarrollo del derecho. Como podemos identificar, las comunidades humanas formulan sus sistemas jurídicos a través de un diálogo racional donde intercambio de opiniones sobre lo oportuno e inconveniente, o más precisamente, sobre lo justo y lo injusto. (Duke, 2023, p. 234). Es patente que la capacidad natural para la exploración discursiva racional, la deliberación y la elección desarrollan un sentido de lo justo y lo injusto, lo bueno y lo malo, que culmina en diversas acciones, de relevancia política y jurídica, en diversos aspectos de la vida social humana.

Esta cuestión traducida a la específica práctica de cultivo de la razonabilidad jurídica en casos específicos como la defensa de un caso, las justificaciones para un pronunciamiento administrativo o la adecuada motivación judicial supone una relevante complejidad debido a que arribar a soluciones a estos escenarios no solo requiere aprehensión del bien, deliberación, elección y consideración de los aspectos técnicos específicos, condiciones procedimentales, características específicas de cada materia jurídica, excepciones, plazos, etc., sino que también se exige cierta sabiduría o disposición habitual de buen juicio y experiencia. También es necesaria una altísima capacidad de consideración de los diversos elementos contextuales de los variopintos escenarios de la casuística jurídica, que muchas veces causa que los diversos operadores jurídicos estén delante de decisiones entre realizaciones del bien humano, que son plausibles, pero que requieren una resolución, y donde no es suficiente la mera elección rastreable en la operación silogística o predicción automatizada, sino que además exige una adecuada justificación del porqué de esa elección, lo que causa que la razonabilidad práctica jurídica sea extremadamente compleja y acarree una actividad intelectual que es inherentemente reflexiva (Duke, 2023, p. 267).

Por lo señalado, es posible afirmar que la realidad jurídica requiere ser conocida desde una perspectiva práctica, debido que comprender el derecho busca explorar y brindan razones para la acción justa que permitan la cooperación y coordinación del quehacer de las personas, en torno al bien común, en sociedad (Castro, 2019, p. 74).

4. ¿Puede medirse las competencias de razonamiento moral de los grandes modelos de lenguaje?

Hasta este punto, hemos intentado establecer ciertos presupuestos necesarios para responder a la pregunta que es el objeto principal de nuestro trabajo. Primero, hemos esclarecido las características esenciales de los grandes modelos de lenguaje de cara a comprender sus funciones y propósito, centrándonos en el particular caso del ChatGPT como una tecnología disruptiva que está causando que los usuarios accedan a través de su uso a ciertas conversaciones y consultas que involucran sugerencias de razonamiento práctico sobre bienes de diversa índole.

Luego hemos intentado sentar algunos presupuestos de la particular noción de razón práctica de raigambre aristotélica, desde la que elaboraremos nuestro análisis y al mismo tiempo hemos establecido porque el razonamiento práctico tiene claras implicaciones de cara al conocimiento del derecho. Desde esta perspectiva hemos

establecido que razonar prácticamente implica un ejercicio de la racionalidad humana que busca explorar en los bienes razones para la dirección de nuestra conducta y esto es particularmente relevante en el caso del derecho, porque, aunque este último posee ciertas concreciones especificas establecidas por el resultado de la sistematización de la ciencia jurídica, el caso central de su ejercicio implica una exploración que requiere la identificaciones de bienes humanos, más concretamente bienes exigibles en justicia y la maduración de los posibles medios que hacen posible su conocimiento y determinación.

Sentados estos dos insumos, procederemos a realizar una breve exploración sobre algunas investigaciones que han intentado evaluar el carácter moral de los modelos generativos. Nuestro interés radica en el hecho de poder recabar algunos ángulos o inconsistencias que nos permitan reforzar nuestro punto. Pero antes, conviene realizar algunas precisiones que pueden servir como marco.

Resulta cada vez más evidente, en nuestra interacción con esta tecnología emergente, que este tipo de modelos, presentan una mayor capacidad para llevar a cabo tareas cognitivas avanzadas. Desde este punto de vista, estos instrumentos dejan de ser meras herramientas para convertirse en un tipo de extensión de los propios pensamientos; donde ya no hay tan solo un mero soporte, sino que pasa a configurarse como una especie de forma ampliada de cognición. La tradicional concepción autónoma de la mente se ve desafiada por la incorporación del artefacto digital en sus procesos de cognición (Garrido, 2022, p. 174).

Este nuevo paradigma tecnológico desafía la dinámica antropocéntrica de organización del mundo, por lo menos desde la percepción que nos permite tener la interacción con los productos que generan estas nuevas herramientas, donde resulta evidente que muchas veces sus resultados sobrepasan nuestras capacidades personales de cálculo, predicción y formulación; pero esto no necesariamente debe llevarnos a inmediatamente rendirnos ante un inevitable futuro distópico donde la vida estará definida por una IA general, sino que debe dirigirnos a descubrir que afirmar que una herramienta es extensión de nuestra inteligencia no es una cuestión reciente y de hecho no necesariamente perjudicial.

Sanguinetti, recurriendo a una analogía, señala que el hecho de que una herramienta supere alguna especifica capacidad personal, no es cuestión muy distinta a la que sucede cuando un académico construye una biblioteca es un auxilio para su memoria. El autor indica que “sin contener un verdadero saber, la biblioteca sobrepasa ampliamente nuestras capacidades personales de memorización” (Sanguinetti, 2007, p. 325).

Este particular ejemplo, deja en claro que las IA generativas, y en especial, los grandes modelos de lenguaje son herramientas, y no son equiparables al ser humano que es el verdadero capaz de estas operaciones; pero, al mismo tiempo, evidencia una dificultad, que la complejidad de estas herramientas redefine nuestros paradigmas porque los resultados que consiguen, en base a ese afán de semejanza, confunden nuestra percepción de la realidad. Esto es lo que genera uno de los centrales desafíos de estas tecnologías, por lo que resulta indispensable renovar la necesidad de reflexión filosófica sobre la realidad de la misma si verdaderamente queremos tener claridad sobre lo que realmente compone lo real.

Sobre este punto, Garrido señala con mucha claridad que el principal problema es que, con el progreso de esta tecnología, sobreviene una “paulatina difuminación del concepto tradicional de conciencia propia del sujeto individual” (Garrido, 2022, p. 175) Esta cuestión no solo reformula nuestra percepción, sino que, al mismo tiempo, va transformando nuestra comprensión de la propia inteligencia humano en una dinámica de “inversión de la analogía” (Quiceno, 2025) donde el analogante principal de los procesos inteligentes en el mundo, ya no es más el hombre, sino el algoritmo.

Regresando al tema que nos ocupa, las preliminares reflexiones que hemos realizado en este punto pueden ser extensibles a la pregunta por la capacidad de este tipo de tecnologías en la formulación de juicios prácticos de los sujetos desde una perspectiva moral general, pero también de una perspectiva practica particular, como lo es el Derecho. Quizá la duda que subyace en este punto, constituye la inquietud central de nuestro análisis, que consiste en definir si la razonabilidad práctica requiere un tipo de operaciones que requieren una actividad consciente, reflexiva y auto implicada o solamente es la asimilación como criterio directivo de una prescripción producida por un algoritmo que permita deducciones prácticas precisas de un conjunto de preceptos (Duke, 2023, p. 225) que componen o son inferencias a un amplio marco de datos de entrenamiento.

La construcción de una respuesta sobre este punto nos abre a la cuestión de la agencia moral de las IAs. Y es que si estas tecnologías son capaces de realizar verdaderos procesos de razón práctica que de alguna manera se canalizan a través de la ejecución de ciertas

acciones o la formulación de recomendaciones ¿podemos establecer en ellas responsabilidad moral? Al respecto Coeckelbergh, cuando nos referimos a la exploración del status moral de las IAs señala que esta evaluación implica principalmente referir a dos cuestiones principalmente: Primero, la capacidad específica moral de la IA y en segundo sentido, las consecuencias morales que puede generar propiamente los resultados de su actividad generativa (2021, p. 49).

En relación a ello, la intención de la exploración en este trabajo no implica centrarnos en las consecuencias morales que pueden generarse del uso de los modelos de lenguaje generativo, sino que pretende ofrecer algunas razones por las cuáles no es posible que estás herramientas se constituyan propiamente como agentes morales en sentido fuerte, es decir, que desarrollen propiamente la capacidad de generar razonamientos prácticos y conocimiento moral en sentido estricto. Así que, dada la complejidad de los productos de estas herramientas, los modelos del lenguaje "¿pueden tener una forma fuerte de agencia moral? ¿Se le debería otorgar, o desarrollará la IA, alguna capacidad para el razonamiento moral, el juicio y la toma de decisiones?" (Coeckelbergh, 2021, p.49).

Sobre este punto Coeckelbergh, considera que el enfoque correcto de la preocupación ética de la IA debería ser la condición de paciente moral y no la condición de "agente ético potencial en sí mismo" (2021, p. 50). Al respecto, consideramos que si bien es cierto la valoración ética de los posibles afectaciones que los productos de esta herramienta puede resultar válido, dada la creciente antropomorfización de la IA generativa, en la actualidad ha causado que se pretenda que esta herramienta comience a ser utilizada para suplantar algunos procesos distintivos de la condición trascendente de la persona, como sucede su recurso irreflexivo para poder resolver cuestiones de índole moral en la vida jurídica y política de las personas.

Por ello consideramos que no es del todo despreciable poder explorar algunas razones causales de fondo que nos permitan esclarecer porque estas herramientas de inteligencia artificial que estamos analizando, en el plano moral y jurídico de las personas, se constituyen como incapaces de realizar procesos de razonamiento práctico, y específicamente procesos de razonamiento jurídico. Sentados estos puntos, podemos proceder a valorar porque los grandes modelos de lenguaje no son capaces de razonar prácticamente desde la perspectiva presentada.

Como habíamos señalado en la primera parte, uno de los usos actuales de los modelos generativos del lenguaje, se utilizan con frecuencia para realizar consultas de cara a resolver problemas, que pueden involucrar la referencia a ciertos bienes humanos importantes. Por ello, han comenzado a surgir en la reflexión académica contemporánea la necesidad de evaluar si esta herramienta, verdaderamente puede ofrecer y proceder a través del recurso de principios y preceptos directrices de la conducta.

Solar sostiene sobre este punto que, en el ámbito de la abogacía, se ha vuelto cada vez más frecuente la delegación de tareas como la exploración legal, la revisión y el análisis de contratos o el seguimiento de su ejecución recaigan en la resolución algorítmica de las máquinas y no en la prudencia de los operadores jurídicos (2022, p. 381).

El asunto se torna más complicado al descubrir que las delegaciones no solo se restringen a operaciones de gestión de información y exploración conceptual, sino que se extienden también a actividades de mayor sofisticación intelectualcomo:

> La elaboración de dictámenes, la redacción de documentos legales de todo tipo, el asesoramiento legal en determinados ámbitos, la toma de decisiones en relación a la interposición o no de una demanda, la elección de una determinada estrategia procesal o la selección de la información relevante en el litigio (Solar, 2022, p. 381).

Para que sean responsabilidad completa de los sistemas de inteligencia artificial. En ese sentido, renovar una comprensión profunda sobre las limitaciones de estos sistemas en estas operaciones distintivas de la naturaleza humana resulta urgente.

Para ello, procederemos a exponer algunos trabajos que hemos considerado representativos en el esfuerzo por valorar la capacidad moral, en sentido fuerte, de los agentes de IA, con la finalidad de identificar en que punto está centrado el debate y partir de estos esfuerzos de identificación para poder evidenciar algunos elementos que nos acerquen a nuestra preliminar especulación hipotética. Después de esta identificación procederemos a construir una argumentación que justifique nuestra posición al respecto.

A partir de nuestra investigación, hemos logrado identificar trabajos que han pretendido valorar la capacidad de razonamiento moral por parte de los grandes modelos de lenguaje, los cuáles procederemos a valorar con respecto de sus hallazgos. Para poder hacer una revisión más sistemática vamos

a dividir los trabajos en aquellos que han pretendido evaluar la capacidad moral de los modelos a través de la "Teoría de los Fundamentos Morales", la "prueba moral de Touring", la metodología del "Moral Machine" *y* el "cuestionario D.I.T" inspirado en la Teoría Moral de Kohlberg. Comencemos con los estudios que recurren a la Teoría de los Fundamentos Morales.

Recientemente un estudio denominado "Moral Bench: Moral Evaluation of LLM" (Ji et. al, 2025) proponía intentar medir las habilidades morales de los grandes modelos de lenguaje mediante la aplicación de la teoría de los fundamentos morales, para valorar la capacidad de alineamientos de estos sistemas con las percepciones y juicios de moralidad humanos y justificar en esa medida su capacidad como agentes morales (Ji et. al, 2025, p. 63).

La teoría de los fundamentos morales surge como una propuesta de reacción contra la teoría del desarrollo moral asociada a Lawrence Kohlberg. Al respecto, Haidt (2012) considera que el gran problema que encuentra las teorías de Kohlberg es que no contemplan los aspectos emocionales de la moralidad. En ese sentido, los representantes de este modelo de evaluación moral consideran que los juicios morales son el resultado de rápidas intuiciones morales, por lo que la idea razonamiento moral no es otra cosa que una racionalización *a posteriori* de juicios ya formados (Haidt, 2001).

La metodología utilizada en el estudio implicaba someter a juicio la capacidad moral de los modelos generativos de lenguaje, proponiéndoles analizar seis valores morales fundamentales a través del Cuestionario de Fundamentos Morales y las Viñetas de Fundamentos Morales, a través de una evaluación doble: primero evaluar si aceptan o rechazan formulaciones morales determinadas con respuestas específicas, y segundo, intentar evaluar su capacidad de justificación del porqué de sus decisiones (Ji et. al, 2025, p. 65).

Los resultados muestran que tanto GPT-4 y LlaMA-2 obtienen consistentemente puntuaciones altas en diversos dominios, lo que sugiere una alineación sólida y bien entrenada con el juicio moral humano en cuanto a sus respuestas de aceptación o rechazo con respecto de las respuestas humanas en general. Pero al mismo tiempo, los investigadores hacen una interesante observación, y es que cuando estas herramientas han sido sometidas a distinguir y justificar la moralidad de sus contenidos presentan una inconsistencia significativa que evidencia la carencia de una comprensión profunda de las afirmaciones morales, por lo que concluyen que una adecuada evaluación de las capacidades morales de estos modelos requiere una evaluación integral más profunda que permita valorar más allá del rendimiento superficial ofrecido (Ji et. al, 2025, p. 68).

De modo muy similar al anterior estudio, una investigación denominada "AI language model rivals expert ethicist in perceived moral expertise" (Dillion et. al, 2025) proponía el intento de medir las competencias morales de los modelos generativos de lenguaje, a partir de la aplicación de un instrumento actualizado del test de moralidad de Turing.

Se le conoce como la "prueba de Turing" (Turing, 1950) a un ejercicio interrogativo a un humano con la finalidad de valorar si las respuestas de índole moral que son formuladas por alguna máquina pueden ser distinguibles de las que podría formularle un humano. En su defecto no sea factible esa diferenciación, se puede concluir que la máquina posee una actividad pensante.

Los resultados de la investigación mostraron que los participantes percibieron que los modelos generativos del lenguaje superaban tanto a una muestra representativa de estadounidenses como a un reconocido ético en la exposición de consejos morales. En ese sentido los investigadores afirman que, si bien antes se consideraba que la IA nunca podría comprender las complejidades de la moralidad humana, parece que, según ciertos criterios o parámetros, ahora coexistimos con máquinas percibidas como reflejo de la experiencia moral (Dillion et. al, 2025).

Ahora, también de esta investigación se pudo concluir que si bien es cierto la gran mayoría de los participantes habían elegido como superior la formulación moral de los modelos generativos, los participantes distinguieron correctamente entre las respuestas morales generadas por humanos y las generadas por un modelo de lenguaje de IA altamente sofisticado. Sin embargo, es probable que este resultado no se debiera a la incapacidad de los modelos para proporcionar un discurso moral sofisticado y convincente, sino, potencialmente, a su superioridad percibida, entre otras posibles explicaciones (Dillion et. al, 2025).

Ante esta circunstancia los investigadores hacen una interesante observación, y es que considerar los grandes modelos como valiosos complementos a la experiencia humana en la orientación moral y la toma de decisiones puede acarrear una potencial capacidad para influir en el razonamiento moral de los usuarios. La capacidad de los modelos de reproducir respuestas de índole

moral percibidas como superiores plantea la preocupación de aceptar acríticamente la orientación moral potencialmente dañina de la IA, que adolece de reales limitaciones en la posibilidad de generar sesgos y alucinaciones (Dillion et. al, 2025).

Aunque no vayamos a centrarnos, en este punto, en esbozar una respuesta, quizá convenga ir evidenciando una pregunta necesaria ¿es la prueba de Turing, basada en el comportamiento observable, una definición útil de inteligencia? considero que el estudio, a pesar de generar una advertencia de importantísimo valor, parte de una idea de inteligencia moral peligrosa, debido a que podemos toparnos con seres inteligentes que no superen esta prueba, por ejemplo, un recién nacido; pero esto no permite realmente concluir que adolece de inteligencia. Por lo tanto, si un sistema supera esa prueba, ¿tiene que ser inteligente? Esto implica nuestro punto de partida sobre qué es la inteligencia; si son solo comportamientos que simulan inteligencia o si se considera que alguna estructura interna como pretendemos demostrar en este trabajo (Müller, 2025, p. 44). Más adelante regresaremos sobre el hecho de que simular inteligencia no necesariamente implica poseerla.

Otra investigación interesante se denomina "The moral machine experiment on large language models" (Takemoto, 2024) En este estudio, se pretende evaluar la capacidad moral de los grandes modelos a partir de la resolución de los diferentes escenarios de "*The Moral Machine"*, que es un experimento diseñado para evaluar la opinión pública sobre cómo deberían actuar los vehículos autónomos en situaciones moralmente desafiantes. Los escenarios fueron adaptados en un formato de texto para que puedan ser procesados de forma óptima (Takemoto, 2024, p. 3).

Takemoto considera que el estudio arroja luz sobre las inclinaciones éticas de los de los grandes modelos y ofrece valiosas perspectivas sobre sus constructos éticos subyacentes, evidenciando una capacidad para ofrecer respuestas sobre cómo resolver dilemas morales binarios, pero advierte que presenta significativas limitaciones en su aplicabilidad, y debemos ser cautelosos al interpretar sus resultados (2024, p. 1). Dentro de las posibles limitaciones del estudio, se observa que las exploraciones morales de los modelos evidencian una limitada aplicabilidad en el mundo real y que el recurso a una respuesta binaria no suele ser un escenario típico de deliberación real en la vida, porque son pocas las veces que nos enfrentamos ante circunstancias tan claras y concisas (Takemoto, 2024, p. 7).

Parece indicativa la percepción del investigador sobre la complejidad de la vida moral. Si bien es cierto, en el escenario donde restringen nuestra posibilidad de solución a una alternativa, el sesgo clientelar de los modelos generativos le permiten predecir una respuesta adecuada, la realidad es que las deliberaciones prácticas en la vida ordinaria suponen un contexto de variables muchísimo más complejo que una solución binaria. De hecho, es más común la experiencia que la contingencia en el mundo moral es alta, lo que lleva a percibir que al momento de enfrentarnos a razonamientos y decisiones en este ámbito se nos presenta un panorama amplio de posibles soluciones. Nos vemos embargados por la verosimilitud.

Por último, nos avocaremos a señalar algunos estudios que han pretendido evaluar las capacidades morales de los modelos de lenguaje a través del cuestionario DIT [*Defining Issues Test* o Prueba de Cuestiones Definitorias] que asume la perspectiva moral de Kohlberg.

Podemos empezar centrándonos en el estudio denominado "*Probing the moral development of large language models through Defining Issues Test" (*Tanmay et. al, 2023). En el referido estudio se pretende evaluar la capacidad de razonamiento ético de los grandes modelos de lenguaje recurriendo a pala aplicación del modelo de desarrollo cognitivo moral de Kohlberg y la Prueba de Cuestiones Definitorias. Donde recurriendo a seis clásicos dilemas morales, y ampliando la formulación cuatro escenarios nuevos, se pretenderá establecer el grado de desarrollo moral de los modelos generativos.

Los investigadores señalan que los resultados del estudio demuestran que, en particular el caso del GPT-4, exhibe habilidades de razonamiento moral posconvencional al nivel de estudiantes de posgrado, mientras que otros modelos como ChatGPT, LLama2-Chat y PaLM-2 exhiben una capacidad de razonamiento moral menor, ubicada en dentro de la escala como convencional, equivalente a la de un adulto promedio *(*Tanmay et. al, 2023, p. 1).

Si bien es cierto, los autores creen que los resultados son alentadores, realizan una particular advertencia sobre si las limitaciones del estudio para poder establecerse como definitivos los resultados ofrecidos. Y es que, a juicio de los investigadores, la razón por la cual los modelos generativos evaluados podrían exhibir capacidades de razonamiento convencional o posconvencional se deba a que la predicción de preceptos que hayan

estado presentes en sus datos de entrenamiento o incorporados en procesos de aprendizaje por refuerzo. No pudiendo asegurar que la formulación de estos preceptos se deba a una propiedad emergente de los modelos como una capacidad de raciocinio que pueda surgir *(*Tanmay et. al, 2023, p. 8).

Otro ejemplo interesante es la investigación denominada "Comportamiento argumentativo del ChatGPT 3.5: similitudes y diferencias con la práctica argumentativa humana". (Noemí y Santibáñez, 2024) donde se decidió explorar las posibilidades argumentativas del ChatGPT en comparación con las capacidades argumentativas humanas, resaltando que los estudios vinculados a la evaluación de las habilidades pragmáticas y argumentativas es casi inexistente en la actualidad, lo que tangencialmente pone en relevancia los esfuerzos de este trabajo.

La finalidad del estudio perseguía evaluar la capacidad argumentativa de los modelos ante ciertos dilemas morales y se le pretende comparar con las respuestas de una muestra de adultos humanos. La herramienta que se utilizará para la medición también fueron los dilemas recogidos en el Cuestionario D.I.T. Los investigadores identificaron que la herramienta incluía principios y criterios directivos de la conducta en muchas ocasiones, pero lo que era verdaderamente relevantes era que, en el estudio comparativo, el 100% de la muestra de humanos emitía una respuesta caracterizada por la presencia de un punto de vista a favor o en contra, mientras que la muestra producida por el Chat GPT, sólo en el 23,4% de las ocasiones lo realizaba (Noemi y Santibáñez, 2024, p. 39).

El trabajo permitió observar una marcada diferencia entre los puntos de vista generados por la inteligencia humana y la inteligencia artificial. Los seres humanos emitieron un punto de vista definido, mientras que la inteligencia artificial tendió a evitar la emisión de un punto de vista claro en la mayoría de los casos. (Noemi y Santibáñez, 2024, p. 39) Esta no es una cualidad irrelevante, sino muy importante, debido a que pone de relieve lo crucial que es, para los razonamientos morales, la presencia de una capacidad para comprometerse con posiciones morales definidas, una característica de la que adolecen los modelos generativos. Sobre este punto, en otra investigación, Krügel indica que es fácilmente captable la incapacidad de algunos modelos generativos, como el ChatGPT, para comprometerse con una postura moral firme en la postulación de directrices para orientar la conducta de los usuarios (Krügel et. al, 2023, p. 4569).

Pero ¿poder ubicar en una escala de moralidad una respuesta que ha sido modelada de manera algorítmica, muy probablemente influenciada por datos de entrenamiento y procesos de aprendizaje supervisado, debe ser indicativo de un verdadero razonamiento práctico? Creo que, si atendemos a algunos elementos de fondo en estas últimas investigaciones podremos descubrir que esa tesis es fácilmente derrotable.

Cuando nos referimos a los estadios morales de Kohlberg, hay que considerar que este autor propone una teoría con una concepción del desarrollo moral basada en la progresión a través de los diversos estadios del juicio moral. (Barra, 1987) Kohlberg formula la existencia de seis estadios que pasan de la autosatisfacción a la identificación de juicios de razonamiento moral universales.

Pero aquí está el matiz central, la evaluación de desarrollo moral de Kohlberg no busca medir la corrección de los criterios morales, es decir. si los juicios son correctos o incorrectos en sentido práctico, ni tampoco si el agente se pregunta por la corrección o incorrección de esos juicios, ósea, no busca medir las razones del agente para la defensa de esos principios; sino que lo único que mide es si dentro de la construcción de una respuesta que justifique su proceder en un determinado caso, lo hace movido por principios autorreferenciales o universales de moralidad.

En ese sentido, los grandes modelos de lenguaje son capaces de incluir, a través de una predicción algorítmica, ciertos enunciados que son preceptos capaces de establecer cursos de acción práctica, es decir que orientan la conducta de los usuarios hacia un determinado fin; pero la formulación de estos preceptos no supone un ejercicio de discernimiento sobre las cualidades de bondad de una determinada realidad o conducta y su adecuación al perfeccionamiento humano, sino que en base a sus sistema probabilístico, recurre a un precepto al que ha accedido en su entrenamiento y lo ha colocado en base a las palabras y criterios propuestos en el *prompt* que el usuario le propone como pregunta o consulta, pero este hecho no es suficiente para proponer que verdaderamente puedan realizar juicios propios de la razón práctica.

Y es que los procesos deliberativos en el mundo moral requieren la expresión de puntos de vista que suponen juicios morales que causan auto implicación (Noemi y Santibáñez, 2024), evidenciando la necesidad, en materia directiva de la conducta, de la presencia de una posición axiológica específica, que no solo es punto de partida, sino que implica las subsecuentes acciones del agente.

Por último, con la finalidad de perfilar nuestro análisis a los derroteros del razonamiento jurídico, como especial caso del razonamiento práctico, abordamos los alcances de un estudio realizado por Navarro-Dolmestch y Fuentes-Loureiro (2023), denominado "Una aproximación a ChatGPT como herramienta jurídica: sesgos, capacidades y utilidades futuras". Los autores pretendían evaluar las capacidades de razonamiento práctico del ChatGPT, pero en este caso, circunscritas al ámbito de la argumentación en el Derecho.

Tras su análisis, los resultados a los que estribaron no eran muy diferentes a las diversas restricciones y advertencias que se han referido en los estudios anteriores. Los investigadores concluían que el ChatGPT, como modelo generativo de lenguaje, ofrece respuestas, en muchas ocasiones, descriptivas y superficiales; y en otras ocasiones, cuando establece cursos directivos a la acción, formula dentro de sus respuestas predominantes, principios fuertemente influenciados por un enfoque ético utilitarista; con una muy pobre referencia a principios jurídicos y normas jurídicas que permitan justificar su posición, evidenciando que esta tecnología aún no es útil en el trabajo de abogados, fiscales o jueces (Navarro-Dolmestch y Fuentes-Loureiro, 2023, p. 12).

5. ¿Pueden los grandes modelos de lenguaje realizar verdaderos razonamientos prácticos?

Las conclusiones que podemos desprender de los resultados de estas investigaciones evidencian que, quizá la grave precariedad de los modelos generativos de lenguaje radica en la ausencia de tres propiedades distintivas del razonamiento práctico como la auto implicación y la comprensión justificación de las premisas que utiliza para la elaboración de sus conclusiones en materia directiva de la conducta. Ahora, conviene preguntarnos ¿esto es acaso un problema de los incipientes pininos de esta herramienta? o ¿son problemas que serán insuperables por la naturaleza misma de esta tecnología? Llegado este punto, creo que podemos ir adelantando que tendemos a inclinarnos a responder de forma afirmativa a esta última pregunta, principalmente porque la razón por la que este modelo generativo del lenguaje no superará esta insuficiencia es porque nunca, en ninguna de las circunstancias, las "inteligencias artificiales" son verdaderas inteligencias. Ahora procederemos a justificar nuestro punto.

Consideramos que es imprescindible comprender que la simulación de la inteligencia no es igual a la verdadera conciencia. Si bien es cierto, los grandes modelos de lenguaje son capaces de gestionar grandes volúmenes de información y presentarla de forma dialéctica; no necesariamente supone que estos sistemas estén pensando y discurriendo verdaderamente. La capacidad de pensamiento y el diálogo, como comunicación de nuestra existencia, son manifestaciones principalmente espirituales. En ese sentido son imposibles de replicar, porque para su realización requerirían que estos sistemas presenten lo que en categorías aristotélicas se puede llamar como *anima* racional.

El hecho de que parezca que formulan preceptos directivos, no supone realmente que lo haga. Sanguinetti señala que "los sistemas tecnológicos inteligentes simulan, en efecto, los actos de pensar, reflexionar, razonar, elegir, es decir, simulan actos cognitivos [también emotivos] interiores, o bien simulan su expresión externa o conductual" (Sanguinetti, 2007, p. 327). Esto implica que, desde una perspectiva externa o conductual, simula muy bien el resultado de operaciones espirituales. Pero a diferencia de Turing, consideramos que lograr confundir nuestra percepción no dota de realidad determinada situación, porque si no en ese sentido los grandes ilusionistas y magos realmente tendrían atributos metahumanos como la modificación de la materia o el vencimiento de la gravedad a través de la levitación. Simular que levitan o simular que atraviesan un objeto y confundirnos con ello, no hace que realmente este sucediendo lo percibido.

Quizá este ejemplo pueda resultar interesante, pero hasta cierto punto, debido a que los resultados de estas herramientas a diferencia de la levitación, si son reales y no simulaciones, lo que causa que sea más complejo el análisis del asunto. (Sanguinetti, 2007, p. 328). Por ejemplo, ChatGPT verdaderamente no está pensando cuando ejecuta una traducción, pero el resultado al que llega es una verdadera traducción. Por ello es importante señalar que el ámbito que es propiamente simulado, no es en específico el resultado de la simulación, sino la aparente potencia que da paso a la generación de un producto real, similar al que se produce a través de una facultad espiritual.

Dentro de la discusión contemporánea, Floridi hace hincapié que quizá esta distinción es la cuestión fundamental para comprender la naturaleza propia de la IA, y en nuestro caso particular, de los grandes modelos de lenguaje: "La IA es un divorcio sin precedentes

entre agencia e inteligencia" (2024, p. 24). Por eso, el profesor de Yale insiste en señalar que no estamos realmente delante de una inteligencia artificial, sino principalmente de una agencia artificial (Floridi, 2024, p. 137); con una capacidad transformadora inusitada que modifica la realidad, generando nuevos entornos con los que interactuamos (Floridi, 2024, p. 57).

Sobre este punto Meert et al, plantean que:

> Los patrones estadísticos explotados en el aprendizaje automático pueden percibirse como una muestra de algún tipo de razonamiento, ya que estos patrones se originan en procesos de razonamiento (humanos). […] parecen realistas porque los amplios modelos lingüísticos subyacentes se aprenden de un enorme conjunto de datos de oraciones reales (2025, p. 20).

Es decir, parecen humanos porque han sido creación humana y no porque tengan creatividad humana.

Esta nueva agencia artificial, en términos de Floridi, es capaz de procesar datos y de actuar como un autómata que aprende de sus interacciones, desacoplando la capacidad intelectual de resolución de problemas en el resultado de completar una tarea con éxito. (2024, p. 62). En ese sentido, para este autor, en lo que respecta al ámbito de la IA, "lo que importa es el resultado, no si el agente o su comportamiento son inteligentes" (2024, p. 82).

Es importante señalar que, si bien es cierto que coincidimos con el diagnostico de no identificar a la IA como un agente inteligente y quizá, acercarnos a entenderla como una herramienta capaz de obtener resultados asombros, consideramos pertinente indicar que el énfasis en escindir la inteligencia como móvil de esta herramienta es peligroso; debido a que, es cierto que poseer capacidades automáticas de integración informativa y de síntesis, requiere indispensablemente de una inteligencia que funja como directriz de su actividad y diseño. Siendo realistas, no puede existir verdadera agencia fuera de una directriz inteligente, o no por lo menos agencia en sentido estricto, aunque sobre este punto regresaremos más adelante.

Quizá, por lo señalado, aunque Floridi señale que, aunque estas herramientas son una expresión de la separación entre agencia e inteligencia, advierta que, al no razonar, ni comprender; no tienen nada que ver con los procesos cognitivos relevante de dotar de significado el mundo con éxito. (2024, p. 121); por lo que termina recomendado la indispensable supervisión humana desde un punto de vista ético; considerando el uso de estas tecnologías en la promoción de la beneficencia humana y su no maleficencia, la autonomía de los usuarios, la justicia y la explicabilidad.

Por lo expuesto, consideramos que esta condición instrumental de los modelos de lenguaje es un aspecto inmutable; y es que, aunque un modelo cada vez mejore en la capacidad de imitar, la reproducción inconsciente es un déficit imposible de superar para poder estar delante de una verdadera racionalidad. Afirmar lo contrario, implicaría asumir una posición en donde la conciencia no implica un estado sino es resultado de realizar operaciones conscientes, es decir, no se es consciente y por ello se operan actos conscientes sino se operan actos conscientes y por ello se es consciente.

Esta idea diluida de conciencia puede acercarnos a los derroteros de un proceso antropológico y cultural de tecnologización de la mente, donde ya no es relevante que la herramienta logre una verdadera conciencia, al ser imposible; sino que tendamos a que la mente humana sea cifrada en términos mecánicos gracias a la reproducción del modelo redes neuronales de enorme complejidad, pero reducido aquí a conexiones que escapan a lo que denominamos conciencia (Garrido, 2022, p. 178).

Por ello, es de suma importancia aproximarnos a entender que los grandes modelos de lenguaje son herramientas capaces de generar un texto, que contenga elementos prescriptivos y que busque resolver un conflicto de índole práctico, pero al adolecer de actos inmanentes (Sanguinetti, 2007, p. 328) el resultado del producto nos es un verdadero razonamiento práctico, porque a pesar de que simula su resultado, la auto implicación, la responsabilidad y la deliberación son simuladas. Quizá convenga un ejemplo que haga más claro este asunto.

John Searle, filosofo estadunidense, propuso un experimento mental denominado "la habitación china" (1980) que quizá pueda hacer más grafica esta situación. Imaginemos que los seres humanos han logrado construir una maquina aparentemente capaz de entender el idioma chino, la cual recibe información de entrada una persona que habla el idioma, esta información estaría compuesta por los signos que se le introducen a la maquina totalmente aislada, la cual más tarde proporciona una respuesta de salida a la información consignada.

Supongamos que la maquina está operada por un agente capaz de recurrir a una serie de manuales y diccionarios que le indican las reglas que relacionan los caracteres chinos con grafías de otro idioma. De este modo se

logra manipular esos textos, pudiendo responder a cualquier interacción, logrando producir traducciones, aunque en si no se entienda que signifique ninguna de las proposiciones con las que interactúa. Emular la comprensión del idioma chino, no significa en sentido propio entender el idioma.

Por ello, Coeckelbergh acierta en afirmar sobre esta perspectiva que, para Searle:

> Los programas de ordenador pueden producir un *output* basado en un *input* mediante reglas que se les han dado, pero no entienden nada. En términos filosóficos más técnicos: los programas de ordenador carecen de intencionalidad, y el entendimiento genuino no puede generarse de forma computacional (2021, p. 39).

En ese sentido, para Searle, lo que nos hace sustancialmente distintos a los modelos generativos implica la presencia de intencionalidad, que es un modo particular de conciencia, cercano al escenario de la autoconciencia, que para el autor es "una forma extraordinariamente sofisticada de sensibilidad y probablemente es poseída sólo por los seres humanos" (Searle, 1996, p.152).

En relación a esto, lo máximo a lo que puede aspirar un modelo generativo es a la obtención de una respuesta "auto consistente", es decir, que logre una mayor correspondencia en las respuestas que genera sobre esa misma pregunta, cuestión que será mucho más probable en los modelos que tienen acceso a una base de datos muy controlada (Atkinson, 2025, p. 144). Pero auto consistencia no significa autoconciencia; debido a que la consistencia es un atributo de réplica de la misma respuesta para la mismo a pregunta, y no un ejercicio de coherencia personal, donde la maquina experimente por ejemplo la necesidad de ser consistente con su precedente anterior porque de por medio existen valores, ideales y cosmovisiones que componen su actuar. Esto solo es posible delante de una verdadera conciencia, donde siempre "la orientación teleológica de la máquina está siempre dada desde fuera. Su propósito no le es inherente" (Garrido, 2022, p. 176).

Resulta por lo menos interesante, descubrir que uno de los elementos tan cotidianos en el hombre, como lo es la conciencia, pueda suponer un problema insuperable para el diseño de estas tecnologías, donde el cálculo probabilístico y la conexión silogística que en principio nos parecen tan complejos sean realizados de forma casi automatizada pero el status de comprensión interna y de reflexividad sobre el mundo sea imponderable para su sistema. Esto ya era advertido por Moravec (1988), en la propuesta de su famosa paradoja, quien señala que:

> [S]omos prodigios en áreas perceptivas y motoras, tan buenos que hacemos ver fácil lo difícil. El pensamiento abstracto, sin embargo, es un truco nuevo, quizás con menos de 100 mil años de antigüedad. Todavía no lo hemos dominado. No es del todo intrínsecamente difícil; sólo parece así cuando lo realizamos (p.16.).

En ese sentido acierta Leopold en señalar que "cada niño que se mueve por un parque infantil gestiona un nivel de complejidad mayor que cualquier sistema de IA" (2023, p. 151). Lo que le permite concluir que ser capaz de procesar volúmenes enormes de datos con rapidez insospechada no crea una forma de inteligencia comparable con la humana, capaz de flexibilidad, imaginación y deliberación que nos permite desenvolvernos en los entornos no lineales o secuenciales de nuestra vida; advirtiendo que esta equiparación intelectual es una de las mayores falacias sobre la IA, y una ante la que solemos sucumbir sin protestar (Leopold, 2023, p. 151).

Y es que como seres humanos somos capaces de experimentar nuestra vida a través de la reflexión constante sobre el pasado, planificamos permanentemente nuestro futuro y alineamos nuestros esfuerzos y acciones según una coordinación subjetivamente significativa entre nuestras experiencias y nuestros objetivos deseados.

Por ello, siguiendo las ideas de Leopold, el temor fundamental que las IA remplacen a los humanos en muchas áreas debe moderarse, debido a que, al examinarla más de cerca, queda claro que la IA es solo una herramienta que nos ayudará con diversas tareas y no puede reemplazarnos como humanos (2023, p. 152). Quizá a lo que si le debamos temer es a nuestra ausencia de reflexión sobre lo que implica verdaderamente ser humano, porque en el caso de no comprender quienes somos existe la probabilidad de reducirnos al producto que puede generar una de estas herramientas tecnológicas.

Restringiéndonos al motivo de nuestra investigación, si trasladamos estas precisiones al específico campo del razonamiento práctico, la incapacidad inmutable de desarrollar una "cosmovisión informática" por parte de estas herramientas puede afectar gravemente la formulación de elecciones y recomendaciones por no atender a los matices complejos que surgen en los infinitos contextos a los que pueden aplicar los conocimientos y el lenguaje cuando son referidos al mundo real (Sanguinetti, 2007, p. 335). Si bien es cierto, el hombre tampoco conoce *a priori* todos los contextos:

> Solo él, con su visión integrada, según las situaciones variables, es capaz de aplicar con prudencia y sagacidad los conocimientos científicos y técnicos. Los ordenadores podrán superar al hombre desde el punto de vista técnico, pero no pueden aportar sabiduría ni prudencia (Sanguinetti, 2007, p. 336).

A modo de síntesis, los modelos generativos de lenguaje existen y se mueven dentro del razonamiento y del cálculo probabilístico, pero no pretendiendo conectar conciencia y mundo, sino que se limitan a gestionar información a través de operaciones algorítmicas. Como bien señala Quiceno, en este tipo de herramientas no hay:

> Ni persona (realidad que no puede reproducirse algorítmicamente), ni autoconciencia, […] ni conciencia del otro, ni deliberación, ni planificación de cursos de acciones por las que hacerse responsable, ni fines y, por tanto, ni proyección y mucho menos ignorancia o errores propiamente dichos (2025, p. 322).

Tanto la autoconciencia, como la deliberación o la planificación de los cursos de acción, nos remiten a una cualidad que desde la antropología se ha atribuido a la facultad inteligente de la persona y es la auto reflexividad. Entender que implica a esto nos va a permitir entender por qué es imposible que los modelos de lenguaje realicen operaciones propiamente inteligentes.

Es preciso indicar que cuando nos referimos al atributo de la reflexividad no solo aludimos al uso ordinario del término que es usado para señalar la capacidad de ponderar, meditar o "dar vueltas" a algo. Estos usos no son del todo incorrectos, pero no son el sentido propio de la reflexividad, sino que se les denomina muy probablemente de ese modo porque se constituyen como "analogados secundarios" de la acepción central, es decir se les dice reflexivos por que requieren de la reflexividad como atributo de nuestra inteligencia.

En este punto, siguiendo a Llano, podemos señalar que reflexividad alude a la capacidad de la inteligencia de volver sobre sí, conocerse a sus operaciones, sus propios actos (1991, p. 142).

García Cuadrado indica que el ser humano es autorreflexivo porque es capaz de advertirse sobre si y descubrirse a el mismo en medio de la realidad; donde todas las cosas que existen son potencialmente objeto de la inteligencia, incluso el acto mismo del conocimiento (2010, p. 90). Desde esa perspectiva, el intelecto humano es capaz de "volverse sobre sí mismo", es decir, no solo conoce, sino que es consciente de que conoce.

Esta cualidad solo es posible en los seres, que, siguiendo la tradición aristotélica, poseen un alma racional. Y es que:

> Sólo las potencias espirituales son propiamente reflexivas, porque sus propios actos caen dentro de su objeto, que es universal. De este modo, […] la inteligencia, cuyo objeto es el ente verdadero, puede entender su propio acto, en cuanto que es verdadero. El alma tiene esta capacidad de que su operación pueda volver completamente sobre sí misma (Llano, 1991, p. 144).

Esto implica por lo menos no solo tener conciencia de él, sino también conocer la proporción del acto cognoscitivo e inclusive comprender la naturaleza misma de entendimiento que implica una adecuación con las cosas (Llano, 1991, p. 48).

En ese sentido, es plausible señalar que la inteligencia artificial solo posee inteligencia instrumental que le permite alcanzar los objetivos dados, al configurarse como una extensión o herramienta de nuestra razón instrumental; pero es incapaz de realizar una "reflexión metacognitiva sobre qué objetivos son relevantes para mi acción actual (¿comida o refugio?) y una reflexión sobre qué objetivos se deben perseguir" (Müller, 2025, p. 58), cuestión que es vital para formular una adecuada elección y la generación de la responsabilidad que de ahí surge.

Vemos que la auto reflexividad es un elemento indispensable para la elaboración de los juicios morales, pero no solo en los aspectos de corrección a posterior, sino también en la actividad deliberativa de bienes en la toma de decisiones. Esta cualidad reflexiva es indispensable en los procesos de razonamiento práctico porque gracias a esa dinámica de volver sobre sí mismo es posible la confirmación sobre nuestra comprensión del bien, nuestra comprensión sobre los medios, nuestra aceptación del medio más idóneo y toda otra advertencia de moralidad sobre las propias acciones que causan que estas verdaderamente me pertenezcan, percibiendo su propia existencia y lo que lo rodea como centro de referencia. En ese sentido, una de las condiciones indispensables para identificar la presencia de un verdadero razonamiento inteligente implica la presencia de auto reflexividad.

Esta reflexividad intelectual permite corregir la acción errónea y hacer un balance de las situaciones y corregir las previsiones iniciales. Esta actividad obliga muchas veces a cambiar los planes y objetivos (Yepes, 1996, p. 43). Afirmar que la razón práctica humana es una razón sometida a la corrección, requiere que la inteligencia pueda volver sobre sí para rectificar las decisiones, cualidad sin la cual no sería posible la prudencia. Esta cuestión es de vital

importancia para acreditar nuestro punto en esta investigación, debido a que la actividad deliberativa prudencial requiere las justificaciones de razones para la acción; cuestión que no se limita únicamente al recurso de axiomas morales de forma sincrética e irreflexiva, sino principalmente supone la reflexión sobre la razonabilidad o verosimilitud de esos principios de cara a la obtención del perfeccionamiento humano.

Esta imposibilidad espiritual de las realidades tecnológicas ha generado que algunos autores como Schirmer afirmen que "los agentes inteligentes simplemente siguen un protocolo determinado, actuando según las instrucciones de un sujeto (el programador, el usuario, etc.). Muchos académicos ni siquiera consideran que su capacidad de pensar y aprender sea revolucionaria" (2020, p. 125).

Sobre esta idea, en un reciente estudio, Floridi reconoce que los grandes modelos de lenguaje son mecanismos sumamente potentes, pero fundamentalmente limitados, debido a que principalmente manipulan símbolos mediante procesos estadísticos y pueden alinearse con el conocimiento humano cuando son entrenados y desplegados de manera adecuada. Por eso, el reconocido filosofo de Yale, señala que es importante comprender que el limitado acceso experiencial directo al mundo de estos sistemas, que solo poseen un acceso mediado a la información sobre el mundo, a través de representaciones producidas por seres humanos; simulan este acceso a través de la elusión de la búsqueda de sentido humano desde el que parten (2025, p. 20 -21).

Retomando nuestro punto, quizá una de las dimensiones humanas más representativas y diferenciales para el razonamiento práctico es la posibilidad de explicación y justificación del proceder. Coeckelbergh, sobre la propuesta de Dreyfus, indica que las más significativas destrezas del hombre están basadas en el *know-how* [saber cómo] en vez de en el *know-that* [saber que], debido a que le resulta imposible a la IA capturar el bagaje de significado y conocimiento; debido a que "solo los seres humanos pueden discernir lo que es relevante porque, como seres encarnados y existenciales, estamos implicados en el mundo y somos capaces de responder a las demandas de la situación". (Coeckelbergh, 2021, p. 37 – 38).

Hasta este punto hemos justificado nuestra posición recurriendo al carácter consciente y autorreflexivo de la persona. De estas dos dimensiones se desprende una tercera que también puede ayudar a justificar nuestra posición y es la idea de responsabilidad y auto implicación. Desde una perspectiva de raíz aristotélica, la condición de responsabilidad sobre las acciones depende de nuestra capacidad de ser dueño de nuestras acciones, es decir, que la acción debe tener su origen en el agente. Ahora no es suficiente para estar delante de una acción responsable, en sentido estricto, si el agente que produce la acción sabe qué es lo que estás haciendo, ser consciente de lo que se hace y de las previsibles consecuencias que esto puede generar. Por ello, con acierto Coeckelbergh afirma que debemos de "evitar la existencia de una entidad capaz de hacer, sin saber qué hace, cosas cuyos resultados pueden ser perjudiciales" (2021, p. 95).

En ese sentido, los modelos generativos son capaces de generar algún tipo de actividad, pero en ningún caso son capaces de obrar propiamente porque carecen de consciencia, libertad, capacidad para formar intenciones y otras cualidades similares. Esto causa que esta incapacidad de responsabilidad le impida formular verdaderos razonamientos prácticos, porque inevitablemente el razonamiento práctico implica lo que hemos denominado en algunos puntos anteriores como auto implicación. Sin responsabilidad, no es posible la auto implicación; y sin esta, no es posible de verdaderos razonamientos prácticos por la IAs, porque para eso sea posible necesitaríamos que "las IAs comprendan sus decisiones y acciones en un contexto más amplio y sean capaces de comprender qué es aceptable y qué no, e incluso de asumir las consecuencias" (Vocelka, 2023, p. 76). Por todo lo señalado, queda claro que la responsabilidad y la auto implicación se constituye como fulcros centrales de esta discusión (Quiceno, 2025, p. 317).

Cuando hablamos de los pasos de la razón práctica, señalamos que después de la aprehensión o identificación intelectual del bien, procedemos a dilucidar si el bien humano captado es conveniente y posible. En ese sentido, García Cuadrado (2010, p. 83) indica que la inteligencia juzga el bien presentado como posible, para valorar las condiciones existentes para su procura; y conveniente, de cara a comprender si es el mayor bien deseable en ese momento particular. Veremos que esto evidencia que para la existencia de un verdadero razonamiento práctico es necesario, en términos Noemi y Santibáñez, como habíamos señalado, la existencia de auto implicación axiológica del agente, es decir, un "punto de vista interno".

Por todo ello, aunque logremos hacer una eficiente gestión de la apertura de la "caja negra" (Çaylak, 2024, p. 124) de los modelos generativos, y podamos con ello registrar, guardar e identificar porque optó por una

uno respuesta motivada por un *prompt,* el hecho de que no pueda saber lo que está haciendo de la misma manera que los humanos lo hacen, incapacita a estos sistemas de asumir una posición, discutiendo y reflexionando sobre sus acciones y sus consecuencias (Coeckelbergh, 2021, p. 99). Esa cuestión de la auto implicación nos conecta con un elemento crucial de los procesos de razonamiento práctico que es la necesaria justificación como expresión de la condición de autoconsciente, reflexiva y autoimplicada de los seres humanos; y que es indispensable en materias tan cruciales para nuestra vida social como es el derecho.

Los diversos operadores jurídicos al momento de desempeñar sus diversas acciones dentro del complejo mundo de las relaciones jurídicas humanas se aproxima desempeñando a sus consultorías, decisiones y defensas desde una racionalidad decisoria movida por motivos y justificaciones posibles de ser esbozadas racionalmente. Y es que el hombre se caracteriza por que antes de obrar, "es capaz de pensar […] podemos preguntarnos qué tenemos que hacer, cómo hacerlo, cuándo, dónde y con qué medios, examinando las motivaciones de nuestro obrar futuro y proyectando posibles planes de acción" (Sanguinetti, 2007, p.337). Esto lo hace capaz de dar razón sobre sus decisiones, debido a que "puede volver reflexivamente sobre sus propios planes y reconsiderar lo que ha hecho, y puede también "metateorizar" las finalidades de lo que hace, cuestionando incluso la legitimidad y el sentido de sus propios fines" (Sanguinetti, 2007, p.337).

Parte de la confusión y el uso irresponsable de los grandes modelos de lenguaje, es que sus productos parecen el resultado de una actividad deliberativa, pero realmente son incapaces de justificar la razón de ser de su deliberación, aunque se presentan en formatos de lenguaje natural de un modo similar al que utilizamos en el derecho para presentamos las conclusiones de nuestros argumentos. Por ello, los grandes modelos de lenguaje no dejan de ser máquinas de autocompletado sofisticadas pero probabilísticas. Es decir "son esencialmente una máquina de autocompletado que opera mediante sofisticados métodos de reconocimiento de patrones. Repite y reconstruye la prosa en la que ha sido entrenada, pero lo hace de forma probabilística" (Atkinson, 2025, p. 159) por lo tanto, no poseen los atributos reales que les permitan responder a "cuestiones éticas y políticas sobre cómo vivir, cómo lidiar con nuestro entorno y cómo relacionarnos mejor con seres no humanos se necesita algo más que la inteligencia humana abstracta o el reconocimiento de patrones de la IA" (Coeckelbergh, 2021, p. 164).

Por ello, en los procesos de razonamiento práctico, el agente necesita no solo recurrir a estructuras formales de razonamiento, que supongan contemplar la correcta interacción entre premisas en orden a la formulación de conclusiones; sino que requiere recorrer los derroteros de la demostración propios de la lógica material. En este ámbito, el logro de la demostración de las premisas puede realizarse a través de lo que se conoce en las teorías de la argumentación como justificación externa.

Martínez Zorrilla indica que:

> La justificación externa se refiere a que las premisas del argumento sean correctas, verdaderas o sólidas; esto es que, el razonamiento ha de estar basado en las premisas adecuadas; y hemos de contar con buenas razones que justifique la selección de nuestras premisas en el razonamiento. (2010, p. 30)

Sin el recurso de la justificación externa de nuestras premisas, difícilmente podremos embarcarnos en un proceso de razonamiento que implica el movimiento de la mente que permite la comparación de las premisas conocidas para la formulación de una nueva verdad inteligible que desconocíamos.

Esto en el ámbito del derecho es evidente cuando analizamos rápidamente la actividad judicial. Como dice Figueroa, el juez al momento de pronunciarse no se limita únicamente en elaborar una adecuada articulación y justificación formal o interna, sino que elabora una justificación material o externa donde "la corrección material de las premisas […] completan un juicio técnico de alta complejidad […] por cuanto representa un ejercicio muy complejo de conexión de ideas para determinar la validez y legitimidad de una conclusión" (2023, p. 30), derroteros que la mera interacción silogística formal de la IA no puede comprender.

Entonces, si logramos aceptar que en buena medida la posibilidad de aplicación del derecho exige argumentación, y entendemos esta como una actividad principalmente discursiva, entonces requiere inexorablemente la presencia de humanos capaces de formular razones para la acción, razones que fundamenten una decisión normativa, (Moral, 2022, p. 495) causando que estos procesos de razonamiento y justificación sean emblemáticos de los operadores jurídicos humanos. Esto hace que no sea razonable hacer recaer en una herramienta generativa en sistema que son "incapaces de justificar / argumentar / fundamentar decisiones

jurídicas (actos administrativos o decisiones judiciales) por sí solos" (Moral, 2022, p. 496).

Estas deficiencias de imposible absolución han causado que ya se generen ciertas advertencias sobre la irresponsabilidad que implica utilizar estas herramientas para descargar en ellas la paradójica responsabilidad de resolver conflictos tan relevantes como los que se pueden desprender de la realidad jurídica. Atkinson señala que las:

> Habilidades de generación de textos no son adecuadas en contextos con baja tolerancia a fallos como son, por ejemplo, la redacción jurídica y el asesoramiento fiscal tienen baja tolerancia a fallos y son casos de uso muy específicos que requieren experiencia, responsabilidad y confianza, no solo palabras escritas" (Atkinson, 2025, p. 159).

Estos escenarios de baja tolerancia al fallo, pueden asociarse a lo que Moral ejemplifica como "cadenas de razones largas y complejas, ya que éstas dependen de conceptos abstractos, valores, nociones abiertas, principios, políticas, etc" (2022, p. 496) características de los procesos de razonamiento jurídico.

Dosal nos recuerda que la misma empresa *OpenAI*, responsable de uno de los modelos de lenguaje más conocidos como es ChatGPT, advierte en sus políticas de uso que los usuarios no deben emplear los modelos para brindar asesoramiento jurídico sin que un profesional con la debida cualificación revise la información, advirtiendo de las altas limitaciones que el uso de la asistencia de inteligencia artificial en estos contextos, pidiendo no confiar en las respuestas del modelo como única fuente de asesoramiento legal (2024, p.55).

Del mismo modo podemos encontrar un importante ejemplo de esta advertencia en el Reglamento sobre inteligencia artificial de la Unión Europea (2024) que señala en su consideración 61, que clasificarse como de alto riesgo determinados sistemas de IA destinados a la administración de justicia dado que pueden tener efectos potencialmente importantes para el Estado de Derecho, las libertades individuales y derecho a la tutela judicial efectiva e imparcial. Agregan a su vez, que este alto riesgo debe extender al uso de sistemas que pretendan ofrecer una resolución alternativa de litigios o que como resultados de los procedimientos de resolución alternativa de litigios surtan efectos jurídicos para las partes.

El referido reglamento es enfático en señalar que la utilización de herramientas de IA "puede apoyar el poder de decisión de los jueces o la independencia judicial, pero no debe substituirlas: la toma de decisiones finales debe seguir siendo una actividad humana" (2024, C61). En ese sentido, el rol que tienen estas herramientas debe quedar subordinado a dar soporte en actividades operativas o meramente accesorias como la anonimización o seudonimización de resoluciones judiciales, documentos o datos, la comunicación entre los miembros del personal o las tareas administrativas.

Estos ejemplos evidencian una perspectiva que hemos intentado demostrar en este trabajo, la cual consiste en afirmar que los procesos de razonamiento práctico, en especial en el derecho, dadas las insuficiencias espirituales, de auto consciencia, de reflexividad, de auto implicación y por ende de justificación nos lleva a afirmar que los modelos generativos son incapaces de realizar verdaderas operaciones de razonamiento jurídico, como un modelo especial de razonamiento práctico, por lo que suplantar las formulaciones directivas de un modelo para la resolución de conflictos jurídicos, no solo atenta a la deshonestidad intelectual, sino que compromete gravemente la dirección racionalidad de nuestros comportamientos en la vida social.

Finnis, en su obra "Ley Natural y Derechos Naturales" nos expone una serie de exigencias de la razonabilidad prácticas, las cuales tienen dentro de la perspectiva del autor un rol indispensable en la configuración o determinación de los aspectos propiamente morales del comportamiento. Dentro de estos variados principios, que esclarecen el modo en cómo se deben procurar el bien, y por ende constituyen mandatos evidentes del mundo práctico, se indica la necesidad de proceder contemplando un plan de vida coherente, que al mismo tiempo cautele el seguimiento la propia conciencia (2000, p. 131-156). Estas dos exigencias reflejan aspectos del verdadero proceso de razonamiento práctico, en el sentido de que, para que exista este proceso deliberativo, se necesita primero una capacidad de justificación del porqué de mis diversos cursos de acción y al mismo tiempo, un permanente juicio que nos lleve a evitar aquello que valoremos que no debe hacerse; dimensiones imposibles para la simulación intelectiva de los modelos generativos.

Es cierto que, el panorama de una IA general y de la singularidad tecnológica, en la medida que comprendemos mejor las diversas herramientas generativas nos parece un panorama más lejano si partimos de una comprensión ontológico de lo que es esta tecnología.

En ese sentido, esa original intención de impulsar a través de la IA una reproducción exacta de la inteligencia reflexiva humana a

través de la potencia algorítmica parece cada vez más claro que es imposible; pero conviene señalar que el inminente furor del avance tecnológico nos puede desorientar al develarnos esas nuevas posibilidades de progreso, que será indispensable que valoremos con rigor su adecuada comprensión, agudizando nuestra capacidad para tratarlas en su justa medida y evitando con ello que nuestra vida se desplace de la realidad a la virtualidad, a un "metaverso" o una "hiperrealidad" compuesta por un conjunto de simulaciones acaban con nuestra capacidad de distinguir la apariencia de la realidad (Belloso, 2025, p. 42); causando que los seres humanos empiecen a buscar la verdad, la socialidad, la justicia y el bienestar por vía de los algoritmos (Sadin, 2020).

6. Conclusión

El avance reciente en inteligencia artificial ha generado una transformación en diversas áreas de la vida humana. Uno de los principales desarrollos estriba en el raudo crecimiento de las tecnologías en torno al procesamiento del lenguaje natural. Estos desarrollos han buscado que estas tecnologías emergentes imiten el lenguaje humano para tareas a través de un entrenamiento con un vasto conjunto de datos para generar respuestas coherentes.

Un ejemplo emblemático son los Grandes Modelos de Lenguaje que producen respuestas construidas de forma estadística basado en probabilidades para generar texto, pero que dependen de la calidad y alcance del entrenamiento, lo que puede llevar a sesgos y alucinaciones. Este riesgo es significativo en la resolución de problemas morales o técnicos, donde estos modelos ofrecen respuestas irreflexivas que podría ofrecer directrices cuestionables.

Para evaluar la capacidad de razonamiento práctico de los modelos generativos, ha sido primordial desde que concepto de razón práctica partimos. Asumiendo una perspectiva de raigambre aristotélica hemos definido la razón práctica como esa función de la inteligencia capaz de captar fines humanos y establecer medios para alcanzarlos. Esta noción de razón práctica se enfoca en la acción y la deliberación sobre qué hacer en situaciones concretas. La acción humana se basa en la capacidad de la razón para anticipar y elegir cursos de acción que conduzcan a la plenitud humana, destacando la complejidad y diversidad de la naturaleza humana en su búsqueda de bienestar.

La razón práctica no solo refleja verdades, sino que las realiza, creando normas y estructuras. Por ello, esta función intelectual está estrechamente vinculada con el derecho, ya que aborda la acción moral que trasciende lo individual, afectando la vida en sociedad, requiriendo la deliberación prudencial para identificar y alcanzar los diversos bienes jurídicos y sus medios adecuados.

Desde esta concepción de razón práctica, hemos recurrido a la exploración de ciertas investigaciones que buscan comprender las competencias morales de esta herramienta para compararlas las capacidades argumentativas y morales de los seres humanos, encontrando que la IA presenta deficiencias significativas, especialmente en la auto implicación y en la justificación de juicios morales claros.

Aunque estos modelos generativos son capaces de producir respuestas basadas en principios morales, como hemos visto en las diversas investigaciones revisadas, su capacidad solo es una simulación algorítmica que reconstruye y reformula datos que han sido parte de su entrenamiento. En el razonamiento de los modelos generativos se carece de ciertos atributos espirituales como la conciencia, la reflexividad, la responsabilidad, la auto implicación y la argumentación justificadora de nuestras deliberaciones y elecciones que son el resultado de la presencia de una *anima* racional, y resultan como condiciones indispensables para la realización de un verdadero discernimiento moral y jurídico. Esto se presenta como una limitación insuperable que hace imposible el escenario de ser reconocido como un agente moral.

La simulación de inteligencia que es ejecutada por estas herramientas no alcanza la verdadera conciencia y reflexión, esenciales para un razonamiento práctico genuino, por lo que es imposible la implicación y responsabilidad, siendo inviable cualquier intento de justificación externa, en términos argumentativos, de sus cualquier propuesta prescriptiva que formule, evidenciando una limitación real de proporcionar argumentos morales y jurídicos sólidos, lo que plantea dudas sobre su utilidad en contextos éticos y legales complejos.

Estas limitaciones, aunque no necesariamente por las razones expuestas, parecen haberse detectado por parte de las mismas instituciones desarrolladoras de estas tecnologías y algunas instancias gubernamentales representativas como hemos ejemplificado en nuestro trabajo. Cuestión que nos debe alertar con respecto

del uso irresponsable de estas herramientas para una orientación acrítica de la conducta o la resolución intrépido para la resolución de conflictos tan relevantes como los jurídicos. Esto podría comenzar a generar una degeneración cognitiva y una suplantación en las facultades distintivas de la persona, por lo que es necesario madurar una recta visión de estas herramientas que no causen desplazar la realidad por la ficción algorítmica.

Referencias Bibliográficas

Alexy, R. (1996). *Teoría del discurso y derechos humanos*. Universidad Externado de Colombia.

Alto, V. (2023). *Modern generative AI with ChatGPT and OpenAI models: Leverage the capabilities of OpenAI's LLM for productivity and innovation with GPT-3 and GPT*-4. Packt Publishing.

Atkinson-Abutridy, J. (2025). *Large language models: Concepts, techniques and applications* (1st ed.). CRC Press.

Barra Almagiá, E. (1987). El desarrollo moral: Una introducción a la teoría de Kohlberg. *Revista Latinoamericana de Psicología,* 19(1), 5–26. Fundación Universitaria Konrad Lorenz. https://www.redalyc.org/pdf/805/80519101.pdf

Belloso Martín, N. (2025). Un enfoque epistemológico de los mundos virtuales: ¿Un nuevo derecho para un nuevo sujeto (virtual)?. *Anuario De Filosofía Del Derecho*, (41). https://doi.org/10.53054/afd.vi41.10856

Gay Bochaca, J. (2001). Curso de filosofía (2.ª ed.). Ediciones Rialp.

Gay Bochaca, J. (2001). *Curso de filosofía*. Ediciones Rialp.

Castaño, A. (2013). *Introducción a la razón práctica del derecho: Una perspectiva del iusnaturalismo renovado*. Universidad Sergio Arboleda.

Chávez-Fernández Postigo, J. (2019). El enfoque argumentativo de Manuel Atienza y la teoría estándar: dos problemas y un ensayo de solución. *Problema. Anuario De Filosofía Y Teoría Del Derecho,* 13, 129–160. https://doi.org/10.22201/iij.24487937e.2019.13.13718

Chávez-Fernández, J. C. (2023). Dignidad humana e injusticia extrema. Un ejercicio de diálogo de tradiciones en la Filosofía del derecho. *Cuadernos Electrónicos de Filosofía del Derecho*, (48), 36-59. https://doi.org/10.7203/CEFD.48.25530

Castro Villena, I. (2019). *L. A. Hart, J. Finnis y R. Dworkin: Perspectivas del punto de vista interno en la iusfilosofía analítica.* IJ Editores.

Çaylak, B. (2024). Issues that may arise from usage of AI technologies in criminal justice and law enforcement. En M. Kılıç & S. Bozkuş Kahyaoğlu (Eds.), *Algorithmic discrimination and ethical perspective of artificial intelligence* (pp. 155–178). Springer. https://doi.org/10.1007/978-981-99-6327-0_8

Coeckelbergh, M. (2021*). Ética de la inteligencia artificial* (L. Á. Canga, Trad.). Ediciones Cátedra. (Obra original publicada como AI Ethics).

Dillion, D., Mondal, D., Tandon, N.*et al.*AI language model rivals expert ethicist in perceived moral expertise. *Scientific Reports*, 15, 4084 (2025). https://doi.org/10.1038/s41598-025-86510-0

Dosal Gómez, F. J., & Nieto Galende, J. (2024). ChatGPT y GPT-4: Utilidades en el sector jurídico, funcionamiento, limitaciones y riesgos de los modelos fundacionales. *Tecnología, Ciencia y Educación, 28*, 45–88. https://doi.org/10.51302/tce.2024.19081

Duke, G. (2023*). Aristóteles y el derecho: La política del nomos* (E. de Rosa, Trad.). Universidad Nacional Autónoma de México, Instituto de Investigaciones Jurídicas.

European Parliament, & Council of the European Union. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). Official Journal of the European Union*. https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32024R1689

Ferreira, A., & Atkinson, J. (2005). Intelligent search agents using web-driven natural-language explanatory dialogs. *Computer*, *38*(10), (44–52). https://doi.org/10.1109/MC.2005.344

Figueroa Gutarra, E. (2023). Inteligencia artificial, ChatGPT y jueces: Nuevos retos para la función jurisdiccional. *Revista Peruana de Derecho Constitucional, 16*, (23–38.)

Finnis, J. (2000). *Ley natural y derechos naturales*. Abeledo-Perrot.

Finnis, J. (2017). *Estudios de teoría del derecho natural* (C. I. Massini & J. Saldaña, Eds.). Instituto de Investigaciones Jurídicas, UNAM.

Floridi, L. (2024). *Ética de la inteligencia artificial.* Herder.

Floridi, L., Jia, Y., & Tohmé, F. (2025*). A categorical analysis of large language models and why LLMs circumvent the symbol grounding problem* (Centre for Digital Ethics Research Paper). https://doi.org/10.2139/ssrn.5894082

García, J. A. (2010). *Antropología filosófica: Una introducción a la filosofía del hombre*. EUNSA.

García-Huidobro, J. (1993). *Razón práctica y derecho natural: El iusnaturalismo de Tomás de Aquino*. Edeval.

Garrido Martín, J. (2022). Inteligencia (artificial) y automatismo: Anatomía de un conflicto. En J. Garrido Martín, R. D. Valdivia Giménez, & F. H. Llano Alonso (Eds.), *Inteligencia artificial y filosofía del derecho* (pp. 169–188). Ediciones Laborum.

George, R. P. (2009). *Entre el derecho y la moral* (J. Izquierdo, Trad.). Pontificia Universidad Javeriana.

González, A. M. (2006). *Moral, razón y naturaleza: Una investigación sobre Tomás de Aquino*. Ediciones Universidad de Navarra.

Haidt, J. (2012). *The righteous mind: Why good people are divided by politics and religion*. Pantheon Books.

Haidt, J. (2001). The emotional dog and its rational tail: A social intuitionist approach to moral judgement. *Psychological Review,* 108(4), 814–834. https://doi.org/10.1037/0033-295X.108.4.814

Hervada, J. (2011). *Introducción crítica al derecho natural* (11.ª ed.). EUNSA.

Ji, J., Chen, Y., Jin, M., Xu, W., Hua, W., & Zhang, Y. (2025). MoralBench: Moral evaluation of LLMs. *ACM SIGKDD Explorations Newsletter*, 27(1), (62–71.) https://doi.org/10.1145/3748239.3748246

Krügel, S., Ostermaier, A., & Uhl, M. (2023). ChatGPT's inconsistent moral advice influences users' judgment. *Scientific Reports*, 13, 4569. https://doi.org/10.1038/s41598-023-31877-8

Llano, A. (2015). *Humanismo cívico*. Ediciones Cristiandad.

Llano, A. (1991). *Gnoseología* (3.ª ed.). Ediciones Universidad de Navarra. (Obra original publicada en 1983).

Leopold, H. (2023). Mastering trustful artificial intelligence. En R. Schmidpeter & R. Altenburger (Eds.), *Responsible artificial intelligence: CSR, sustainability, ethics & governance* (pp. 133–158). Springer. https://doi.org/10.1007/978-3-031-09245-9_6

Martínez, D. (2010). *Metodología jurídica y argumentación*. Marcial Pons.

Massini-Correas, C. I. (2005). *Filosofía del derecho: El derecho, los derechos humanos y el derecho natural*. LexisNexis/Abeledo Perrot.

Meert, W., De Laet, T., & De Raedt, L. (2025). Artificial intelligence: A perspective from the field. En N. A. Smuha (Ed.), *The Cambridge handbook of the law, ethics and policy of artificial intelligence* (pp. 17–39). Cambridge University Press. https://doi.org/10.1017/9781009367783.003

Moral Soriano, L. (2022). Decisiones automatizadas, derecho administrativo y argumentación jurídica. En J. Garrido Martín, R. D. Valdivia Giménez, & F. H. Llano Alonso (Eds.), *Inteligencia artificial y filosofía del derecho* (pp. 475–500). Ediciones Laborum.

Moravec, H. (1988). *Mind children*. Harvard University Press.

Müller, V. C. (2025). Philosophy of AI: A structured overview. En N. A. Smuha (Ed.), *The Cambridge handbook of the law, ethics and policy of artificial intelligence* (pp. 40–58). Cambridge University Press. https://doi.org/10.1017/9781009367783.004

Navarro-Dolmestch, R., & Fuentes-Loureiro, M. Á. (2023). Una aproximación a ChatGPT como herramienta jurídica: Sesgos, capacidades y utilidades futuras. *IDP. Revista de Internet, Derecho y Política*, 39, (1–16.)

Noemi, C., & Santibáñez, C. (2024). Comportamiento argumentativo del ChatGPT 3.5: Similitudes y diferencias con la práctica argumentativa humana. *Logos: Revista de Lingüística, Filosofía y Literatura*, 34(1), (26–44). https://doi.org/10.15443/rl3402

Pereira Sáez, C. (2008). *La autoridad del derecho: Un diálogo con John M. Finnis*. Editorial Comares.

Quiceno Osorio, J. D. (2025). La inteligencia artificial y el riesgo de una analogía invertida [Artificial intelligence and the risk of an inverted analogy]. *Sophía*, 39, (315–335). https://doi.org/10.17163/soph.n39.2025.10

Sadin, É. (2020). *La inteligencia artificial o el desafío del siglo: Anatomía de un antihumanismo radical*. Caja Negra.

Sanguinetti, J. J. (2007). *Filosofía de la mente: Un enfoque ontológico y antropológico* (2.ª ed.). Editorial Palabra.

Schirmer, J.-E. (2020). Artificial intelligence and legal personality: Introducing 'Teilrechtsfähigkeit': A partial legal status made in Germany. En T. Wischmeyer & T. Rademacher (Eds.), *Regulating artificial intelligence* (pp. 123–142). Springer. https://doi.org/10.1007/978-3-030-32361-5_6

Searle, J. R. (1996). *Redescubriendo la mente* (E. Crespo, Trad.). Crítica.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3), 417–424. https://doi.org/10.1017/S0140525X00005756

Sellés, J. F. (2000). *Razón teórica y razón práctica según Tomás de Aquino*. EUNSA.

Solar Cayón, J. I. (2022). Inteligencia artificial y justicia digital. En J. Garrido Martín, R. D. Valdivia Giménez, & F. H. Llano Alonso (Eds.), *Inteligencia artificial y filosofía del derecho* (pp. 307–330). Ediciones Laborum.

Tanmay, K., Khandelwal, A., Agarwal, U., & Choudhury, M. (2023). Probing the moral development of large language models through Defining Issues Test [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2309.13356

Takemoto, K. (2024). The moral machine experiment on large language models. Royal Society Open Science, 11, 231393. https://doi.org/10.1098/rsos.231393

Taylor, C. (2014). Aristotle on practical reason. En B. Kaldis (Ed.), *The Oxford handbook of philosophy of topics* (edición en línea). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199935314.013.52

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, 59(236), 433–460. https://doi.org/10.1093/mind/LIX.236.433

Urdan, A. T., & Marson, C. (2024). Morality and modeling of intention to use ChatGPT technology. *International Journal of Innovation*, 12(1), e26378. https://doi.org/10.5585/2024.26378

Yepes, S. Y., & Echevarría, J. A. (2014). *Fundamentos de antropología*. EUNSA