ISSN-e: 2444-2887

ChatGPT y GPT-4: utilidades en el sector jurídico, funcionamiento, limitaciones y riesgos de los modelos fundacionales

Francisco Julio Dosal Gómez (autor de contacto)

Abogado/Graduado en Derecho por la Universidad de Cantabria/LLM en Derecho Internacional de los Negocios en el Centro de Estudios Garrigues (España)

francisco.julio.dosal@studentsceg.com | https://orcid.org/0009-0006-0506-5120

Judith Nieto Galende

Abogada/Doble grado en Derecho y Administración de Empresas por la Universidad Autónoma de Madrid/ LLM en Derecho Internacional de los Negocios en el Centro de Estudios Garrigues (España)

judng6997@gmail.com | https://orcid.org/0009-0003-8094-4449

Este trabajo ha obtenido un **Accésit** del **Premio Estudios Financieros 2023** en la modalidad de **Educación y Nuevas Tecnologías**. El jurado ha estado compuesto por: D. Alfonso Gutiérrez Martín, D.ª Verónica Marín Díaz, D. Joaquín Paredes Labra, D. Francisco Roca Rodríguez y D. Javier Manuel Valle López. Los trabajos se presentan con seudónimo y la selección se efectúa garantizando el anonimato de los autores.

Extracto

Los sistemas de inteligencia artificial como ChatGPT, el *chatbot* de OpenAI, basado en la familia de modelos de lenguaje GPT (*generative pre-trained transformers*), así como aquellas otras soluciones basadas en esta tecnología y ajustadas para tareas específicas, han despertado un gran interés en diversos ámbitos, entre los que se incluyen el sector legal y, particularmente, el sector de la abogacía. Sin embargo, tales modelos presentan todavía importantes limitaciones y riesgos asociados a su empleo y funcionamiento, que deben ser considerados a fin de hacer un uso adecuado y jurídicamente responsable de esta tecnología. El presente trabajo tiene por objeto aproximar a los lectores (hombres y mujeres) a la configuración, a la arquitectura y al funcionamiento de estos sistemas, así como a sus funcionalidades dentro del sector jurídico, incluyendo una revisión a sus limitaciones y riesgos jurídicos asociados, con importantes implicaciones prácticas en su aplicación.

Palabras clave: ChatGPT; GPT-4; OpenAI; inteligencia artificial; tecnología legal; procesamiento del lenguaje natural; propiedad intelectual; protección de datos; innovación en la industria legal.

Recibido: 03-05-2023 | Aceptado: 08-09-2023 | Publicado (en avance *online*): 15-03-2024

Cómo citar: Dosal Gómez, F. J. y 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

ISSN-e: 2444-2887

ChatGPT and GPT-4: utilities in the legal sector, functioning, limitations and risks of foundational models

Francisco Julio Dosal Gómez (corresponding author)
Abogado/Graduado en Derecho por la Universidad de Cantabria/LLM en Derecho Internacional de los Negocios en el Centro de Estudios Garrigues (España)
francisco.julio.dosal@studentsceg.com | https://orcid.org/0009-0006-0506-5120

Judith Nieto Galende
Abogada/Doble grado en Derecho y Administración de Empresas por la Universidad Autónoma de Madrid/ LLM en Derecho Internacional de los Negocios en el Centro de Estudios Garrigues (España)
judng6997@gmail.com | https://orcid.org/0009-0003-8094-4449

This paper has won a **Runner-up Prize in the Financial Studies 2023 Award** in the category of **Education and New Technologies**. The jury members were: Mr. Alfonso Gutiérrez Martín, Mrs. Verónica Marín Díaz, Mr. Joaquín Paredes Labra, Mr. Francisco Roca Rodríguez and Mr. Javier Manuel Valle López. The entries are submitted under a pseudonym and the selection process guarantees the anonymity of the authors.

Abstract

Artificial intelligence systems such as ChatGPT, the OpenAI chatbot, based on the language model family GPT (generative pre-trained transformers), as well as other solutions built on this technology and fine-tuned for specific tasks, have generated considerable interest across various sectors, including the legal sector. However, such models still feature important limitations and legal risks associated to their use, which must be considered in order to make a proper and legally responsible use of this technology. This work aims to familiarize the readers (men and women) with the configuration, architecture, and functioning of these systems, as well as their functionalities in the legal sector. It includes a review of their associated legal limitations and risks, with crucial practical implications in their application.

Keywords: ChatGPT; GPT-4; OpenAI; artificial intelligence; legal tech; natural language processing; intellectual property; data protection; legal industry innovation.

Received: 03-05-2023 | Accepted: 08-09-2023 | Published (online preview): 15-03-2024

Citation: Dosal Gómez, F. J. and Nieto Galende, J. (2024). ChatGPT and GPT-4: utilities in the legal sector, functioning, limitations and risks of foundational models. *Tecnología, Ciencia y Educación*, *28*, 45-88. https://doi.org/10.51302/tce.2024.19081

Sumario

Nota: los autores del artículo declaran que todos los procedimientos llevados a cabo para la elaboración de este estudio de investigación se han realizado de conformidad con las leyes y directrices institucionales pertinentes.

1. Introducción

Resulta indubitado que el concepto de «inteligencia artificial» ha evolucionado significativamente desde que John McCarty se refiriera al problema de la inteligencia artificial como el consistente «en hacer que una máquina se comporte de un modo que se consideraría inteligente si lo hiciera un ser humano» (McCarthy *et al.*, 1955), constituyendo, en la actualidad, uno de los sectores más disruptivos y revolucionarios. Según la Organización para la Cooperación y el Desarrollo Económicos (OCDE), un «sistema de inteligencia artificial es un sistema basado en máquinas que, por objetivos explícitos o implícitos, infiere, a partir de la entrada que recibe, cómo generar salidas tales como predicciones, contenidos, recomendaciones o decisiones que pueden influir en entornos físicos o virtuales, añadiendo que los distintos sistemas de inteligencia artificial varían en sus niveles de autonomía y capacidad de adaptación tras su despliegue»[1] (OCDE, 2019, p. 7). En palabras de Satya Nadella, CEO de Microsoft, asistimos al inicio de la edad de oro de la inteligencia artificial (World Economic Forum, 2023) y la expectación, tanto privada como pública, por estos nuevos sistemas resulta cada vez más evidente. Algunos estudios han valorado que en 2022 el tamaño del mercado mundial de inteligencia artificial ascendía a 136.550 millones de dólares, con una expansión a una tasa de crecimiento anual compuesto (*compound annual growth rate* [CAGR]) del 37,30 % entre 2023 y 2030 (Grand View Research, 2023); mientras que otros lo situaban en 2022 en 428.000 millones de dólares, con una expansión del 21,60 % entre 2023 y 2030 (Fortune Business Insights, 2023); concluyendo que el mercado de inteligencia artificial alcanzará 2.000 billones de dólares para 2030.

Entre las razones de esta rápida evolución, además de los importantes avances en los diferentes subcampos que informan esta tecnología, como el aprendizaje automático (*machine learning*) (Janiesch *et al.*, 2021), las ciencias de la computación o el crecimiento en el volumen de datos disponibles, se encuentra el desarrollo de los denominados «modelos de lenguaje extenso» (*large language model* [LLM]), los cuales han demostrado un rendimiento notable en diversas tareas de comprensión, procesamiento y generación de lenguaje natural (Brown, 2020; Chowdhery *et al.*, 2022). Los principales exponentes de los LLM actuales –como GPT-3 (Brown, 2020) o GPT-4 (Bubeck *et al.*, 2023), InstructGPT (Ouyang *et al.*, 2022), FLAN (Wei, Bosma, *et al.*, 2022), PaLM (Chowdhery *et al.*, 2022), LLaMA 1 y 2 (Touvron *et al.*, 2023), entre otros ejemplos destacables (Bai *et al.*, 2022; Zeng *et al.*, 2023; Zhang *et al.*, 2022; Xu *et al.*, 2023)–, gracias a un complejo proceso de entrenamiento y

[1] La cita textual ha sido traducida por los autores de este estudio de investigación a partir del documento original en inglés.

capacitación a través del empleo de algoritmos de aprendizaje automático y, en ocasiones, de sistemas de retroalimentación –con y sin intervención humana, en función del modelo–, codifican conocimientos globales dentro de sus parámetros (Han *et al.*, 2021; Huang y Chang, 2023). Tales modelos presentan diferentes grados de comprensión literal, contextual y sentimental de palabras, frases y textos en lenguaje natural, permitiendo la interacción natural y directa con los usuarios. A ello se añade la posibilidad de que los modelos recuperen durante su funcionamiento información empleada durante su entrenamiento, sin perjuicio de ser capaces de aprender y elaborar respuestas en entornos desconocidos –en tareas para las cuales no han sido entrenados específicamente– o a partir de un número reducido de ejemplos o indicaciones previas (Kojima *et al.*, 2022; Pu y Demberg, 2023; Wei, Bosma *et al.*, 2022).

La evolución de estos modelos se encuentra en la base del desarrollo de los sistemas de inteligencia artificial de propósito general y de los controvertidamente denominados como «modelos fundacionales» (*foundation models*) (Bommasani *et al.*, 2022), definidos como sistemas de gran magnitud que son capaces de adaptarse a una amplia variedad de propósitos diferentes –al poder realizar una amplia gama de tareas diferenciadas, como la generación de vídeo, texto e imágenes, la conversión en lenguaje lateral, la informática o la generación de códigos informáticos, incluyendo aptitudes de razonamiento lógico e inferencial, en ocasiones superiores a los humanos, lo que permite a los mismos la deducción de información implícita y la inferencia de conclusiones a través de ciertos procesos lógicos– y que pueden ser implementados como base para el desarrollo de otros sistemas de inteligencia artificial más específicos.

OpenAI, el grupo empresarial estadounidense creador de la familia de modelos de lenguaje extenso GPT (*generative pre-trained transformer*) y de diferentes soluciones basadas en esta tecnología, como el conocido sistema de *chatbot* ChatGPT (*chat generative pre-trained transformer*), ha revolucionado este sector, implementándose por un número cada vez mayor de compañías, ya directamente o a través de otras aplicaciones y servicios, incluido el empleo de aquellas soluciones de interfaces de programación de aplicaciones (*application programming interface* [API]).

Sin embargo, este conjunto diverso de tecnologías plantea, a su vez, importantes riesgos asociados que afectan a diferentes aspectos sociales, éticos, técnicos, económicos y jurídicos, muchos de los cuales son transversales a un amplio número de soluciones basadas en la inteligencia artificial. Su sistematización, dada la amplitud, complejidad y rápida evolución de este fenómeno, unida a las particularidades inherentes a cada modelo –así como a aspectos relevantes del contenido de los datos empleados en su desarrollo, capacitación y entrenamiento– obligará a analizar las circunstancias, los riesgos asociados y el régimen jurídico aplicable a cada situación de hecho por separado. A ello se añade el carácter eminentemente digital e internacional de la prestación de servicios de las empresas propietarias de los modelos, la pluralidad de sujetos y partes involucradas, los flujos de información o el contenido de la interacción por parte de los usuarios y de los resultados

proporcionados por los mismos, incluyendo el empleo particular de dichos resultados y la evolución del marco jurídico aplicable, que habrán de analizarse adecuadamente en orden a su debida implementación en el mercado.

Entre los aspectos jurídicos que han suscitado mayores preocupaciones respecto a la utilización comercial de soluciones basadas en inteligencia artificial, se incluyen aquellos relacionados con la normativa en materia de protección de datos, confidencialidad y seguridad de la información, propiedad intelectual, derechos de autor y propiedad industrial, incluyendo aspectos en materia de derecho de la competencia y competencia desleal, a los que se suman aquellos derivados del cumplimiento de la creciente normativa específica en materia de inteligencia artificial. En combinación con los anteriores se incluyen aquellas preocupaciones por su fiabilidad ante la existencia de importantes limitaciones y sesgos, incluida la posibilidad de alucinaciones y errores en los modelos, a los que se suman sus aspectos éticos y la posibilidad de su uso sin el debido control, lo que podrá derivar en importantes supuestos de responsabilidad contractual o extracontractual, incluyendo, entre otros, posibles infracciones en materia profesional o disciplinaria y que, generalmente, no estarán bajo la cobertura de seguros de responsabilidad civil.

No obstante, a pesar de sus riesgos, tales modelos continúan despertando un creciente interés en una amplia variedad de sectores, especialmente en aquellos dominados por el uso del lenguaje, como el sector jurídico, siendo cada vez mayor el número de firmas nacionales e internacionales que optan por implementar este tipo de soluciones, advirtiendo de un importante cambio de paradigma en el futuro de la profesión jurídica. El presente estudio tiene por objeto aproximar al lector a la estructura y al funcionamiento de modelos como ChatGPT y GPT-4 y a sus funcionalidades en el sector jurídico, incluyendo una revisión a las principales limitaciones y riesgos y a las propuestas más destacadas en la normativa comunitaria.

2. Arquitectura y funcionamiento de GPT-4 y ChatGPT

La adecuada comprensión de las funcionalidades, de las limitaciones y de los riesgos inherentes al empleo de sistemas como ChatGPT o soluciones basadas en modelos como GPT-4 obliga a realizar una primera aproximación a ciertos aspectos relativos a su entrenamiento, estructura y funcionamiento. La versión más reciente de ChatGPT se basa en GPT-4, la actual iteración de la familia de modelos de LLM multimodal de OpenAI, lanzado en marzo de 2023 y que introduce mejoras significativas en términos de rendimiento, capacidad y habilidades respecto a su predecesor GPT-3.5 (OpenAI, 2023c). A causa del panorama competitivo y de las implicaciones de seguridad de los modelos a gran escala, OpenAI no ha revelado los detalles técnicos del modelo, así como tampoco información sobre el mecanismo para la creación del conjunto de datos de entrenamiento, el *hardware* utilizado durante esta fase, el tamaño o el número de parámetros, su arquitectura u otros factores como la

inferencia o el grado de aprendizaje. Sin embargo, diversos autores suponen que las diferencias de GPT-4 frente a sus predecesores son fundamentalmente cualitativas o de escala y que no existen cambios sustanciales de diseño (Bowman, 2023).

A grandes rasgos, el modelo permite el procesamiento simultáneo de diferentes datos de entrada secuenciales por parte de los usuarios, a partir de un *prompt* de entrada redactado en lenguaje natural (Vaswani *et al.* 2017), efectuando una predicción probabilística de la secuencia, *token* por *token*, e identificando cuál será el siguiente *token* más probable, lo que le permitirá ponderar y proporcionar una respuesta de texto de salida en formato de texto o código. A partir del *prompt* de entrada y mediante el empleo de diferentes procesos de descomposición, vectorización y mecanismos como el de «autoatención», el modelo no solo asigna un contexto o significado a cada palabra de la oración, sino que pondera diferencialmente la importancia de cada uno de los elementos incluidos en la secuencia de datos de entrada, su contexto y los datos de su propia respuesta. A tal fin, el modelo ajusta su comportamiento de forma incremental a través de un sistema de decodificación autorregresiva, reajustando su predicción en base al contexto proporcionado por los *tokens* que le preceden en su propia respuesta para determinar la probabilidad de los siguientes en la generación secuencial, hasta completar la misma, lo que ocurrirá cuando las capas del transformador emitan un *token* de parada. Sin embargo, la doctrina actual carece de una comprensión clara sobre cómo funciona en realidad, cuándo y por qué falla en ocasiones, incluyendo aspectos tales como el surgimiento de habilidades emergentes (Bommasani *et al.*, 2022), no existiendo, por el momento –sin perjuicio de que haya ciertos estudios actuales en la materia (Burns *et al.*, 2022; Chan *et al.*, 2022; Elhage *et al.*, 2021; Lovering y Pavlick, 2022)–, ninguna técnica que permita comprender completamente la forma en que el modelo efectúa estas predicciones o el tipo de conocimiento, razonamiento u objetivos subyacentes del modelo cuando produce un resultado (Bowman, 2023).

Las interacciones con LLM, como GPT-4, se benefician de la claridad y de la especificidad del *prompt* empleado por el usuario (White, Fu *et al.*, 2023; White, Hays *et al.*, 2023). Aunque modelos de la familia GPT, como GPT-3 (Brown *et al.*, 2020) o GPT-4, adviertan limitaciones en actividades como la síntesis de texto, el razonamiento o la explicabilidad, incluyendo diversas tareas de procesamiento del lenguaje natural –y sin perjuicio de que otros estudios atribuyan diferentes grados de capacidad a los modelos en configuraciones de ninguna o pocas instancias (*zero-shot* y *few-shot*) (Brown *et al*., 2020; Chalkidis, Fergadiotis, Kotitsas *et al.,* 2020; Kojima *et al.*, 2022; Wei, Bosma *et al*., 2022)–, diversos autores señalan que su capacidad puede aumentar sustancialmente a través de razonamientos paso a paso, ya sea mediante ajuste fino (Cobbe *et al.*, 2021; Nye *et al.*, 2022; Rajani *et al.*, 2019; Zelikman *et al.*, 2022) o la generación de cadenas de pensamiento.

Según diversos estudios, la generación de cadenas de pensamiento (*chain of thoughts*) –esto es, pasos intermedios de razonamiento, al proporcionar ejemplos al modelo o la solicitud de un razonamiento «paso a paso»– mejora significativamente la capacidad de

los LLM para realizar razonamientos complejos (Kojima *et al.*, 2022; Wei, Wang *et al.*, 2022), sin perjuicio de que otros autores hayan propuesto sistemas alternativos como el de autoconsistencia (Wang *et al.*, 2023). En la elaboración de resúmenes de artículos y documentos, el empleo de un *prompting* adecuado permitirá un control preciso de las características del mismo, como su longitud (Goyal *et al.*, 2023), los temas (Bhaskar *et al.*, 2023) y el estilo (Pu y Demberg, 2023). Algunos estudios sugieren la utilidad de la elaboración de cadenas de densidad (*chain of density* [COD]) en las que se genere un resumen inicial que incorpore iterativamente elementos destacados del documento a resumir, generando un resumen mucho más denso, preciso y con menor pérdida de información (Addams *et al.*, 2023). A ello se añade la posibilidad de brindar al modelo instrucciones personalizadas, agregando preferencias o requisitos respecto a la generación de sus respuestas (OpenAI, 2023b) o la parametrización de respuestas en soluciones API, mediante la indicación a partir de parámetros como «*temperature*», «*top_p*», «*stop*», «*best_of*», «*n*», «*max_token*», «*presence_penalty*», «*frequency_penalty*» o «*logit_bias*», entre otros.

El preentrenamiento del modelo de transformador que sirve de base a GPT-4 se fundamenta en un sistema de aprendizaje semisupervisado que incluye una primera fase de aprendizaje no supervisado y una fase posterior de ajuste fino supervisado, afinado mediante un sistema de aprendizaje por refuerzo a partir de la retroalimentación humana (*reinforcement learning from human feedback* [RLHF]) (Fernandes *et al.*, 2023; OpenAI, 2023c; Ouyang *et al.* 2022; Touvron *et al.*, 2023).

En una primera fase, dirigida a recopilar datos de demostración y entrenar una «política» supervisada, se extrae un *prompt* del *dataset* de entrenamiento. Este *prompt* se somete a una evaluación por una persona humana, quien habrá de determinar cuál es el resultado (*output*) que representa el comportamiento más deseado, criterio que se usará para el ajuste fino supervisado (*supervised fine-tuning* [SFT]) (Zhang, Dong *et al.*, 2023). En una segunda fase, dirigida a recoger datos de comparación y entrenar un modelo de recompensa (*reward model* [RM]), se extrae nuevamente un *prompt* y diversos *outputs* calificados como deseados, a fin de clasificarlos de mejor a peor de acuerdo con el criterio de un evaluador humano. Esta información será empleada para el entrenamiento del RM, que permitirá entrenar el modelo para predecir el resultado preferido por los humanos. En una tercera fase, y con el objetivo de optimizar una «política» –esto es, la estrategia que el modelo aprendió a usar para lograr su objetivo– con respecto al modelo de recompensa, se extrae nuevamente un *prompt* al que se aplica un algoritmo de aprendizaje por refuerzo de optimización de política proximal (*proximal policy optimization* [PPO]) (Schulman *et al.*, 2017).

Así, tras la aplicación del PPO, se genera un *output* que será sometido al RM, que calculará la recompensa para el mismo, la cual será empleada a su vez para actualizar la «política» utilizando el PPO, permitiendo su retroalimentación sin intervención humana. A ello se añaden otras herramientas y sistemas de ajuste posteriores, como los modelos de recompensa basados en reglas, a fin de evitar comportamientos indeseados. El objetivo

principal del RLHF es incorporar la experiencia y el conocimiento humanos a los algoritmos de aprendizaje automático para mejorar su rendimiento y su capacidad para resolver tareas complejas.

Sin embargo, uno de los aspectos con mayores implicaciones, como veremos, se basa en el preentrenamiento de la familia de modelos GPT y, particularmente, de GPT-4, a partir del contenido del *dataset* conformado por un corpus de texto no etiquetado de cientos de miles de datos. Según la propia OpenAI, este *dataset* estaría conformado por fuentes de acceso abierto en internet, datos sujetos a licencia de terceros e información creada por usuarios y revisores humanos (OpenAI, 2023c, 2024c). Ahora bien, es importante tener en cuenta que los modelos no copian ni almacenan información de capacitación, como ocurriría con la creación de base de datos al uso, sino que aprenden de la misma, careciendo de un acceso posterior a dicha información de entrenamiento tras haber aprendido de ella. Modelos como GPT-4 aprenderían, a partir de los referidos datos, a formular asociaciones entre palabras, de forma que dicho aprendizaje ayudaría al modelo a actualizar sus parámetros, permitiéndole posteriormente predecir y generar nuevas palabras en respuesta a las solicitudes del usuario (OpenAI, 2023c, 2024c).

Aunque OpenAI no haya revelado demasiada información acerca del *dataset* de preentrenamiento de GPT-4, diversos autores plantean que el modelo pudo haber sido preentrenado con fuentes como Common Crawl, WebText2, Books1 y Books2, así como Wikipedia en inglés, conformando un total de 499.000 millones de *tokens* en 753.4 GB de contenido (Roberts, 2022; Thompson, 2022), al que se añadiría un conjunto de datos con licencia, autorizados por terceros proveedores sin identificar. Gran parte de los referidos corpus se confeccionan mediante técnicas de raspado web (*web scraping*), extrayendo y recopilando datos de forma indiscriminada de miles de páginas web, posteriormente empleadas para la capacitación de los modelos, lo que plantea, como veremos, graves riesgos jurídicos.

3. ChatGPT con GPT-4: tendencias en el sector jurídico y en el sector de la abogacía

La evolución experimentada en los sistemas de inteligencia artificial, especialmente de aquellos sistemas basados en el lenguaje natural, permite el desarrollo de soluciones aplicables a una amplísima variedad de actividades, incluyendo aquellas que podrían calificarse como propiamente jurídicas.

La incidencia de los sistemas de inteligencia artificial en el sector jurídico alcanza aspectos de enorme diversidad, tanto en el sector privado como en el ámbito estrictamente público. En España, el Real Decreto-Ley 6/2023, de 19 de diciembre, por el que se aprueban medidas urgentes para la ejecución del plan de recuperación, transformación y resiliencia en materia de servicio público de justicia, función pública, régimen local y mecenazgo, ha

introducido el principio de «orientación al dato» en la Administración de Justicia, previendo la incorporación de sistemas de inteligencia artificial para apoyar la función jurisdiccional.

También la Administración Tributaria, tras el empleo de sistemas informáticos de selección, obtención y tratamiento automático de la información, comienza a añadir sistemas basados en inteligencia artificial, tanto para actividades de información y asistencia (Agencia Tributaria, 2020) como, paulatinamente, para actividades de investigación frente al fraude fiscal y el blanqueo de capitales, lo que plantea importantes cuestiones sobre límites y derechos de los contribuyentes (Rincón, 2023).

Asimismo, en el ámbito laboral, se observa la creciente implementación de sistemas de inteligencia artificial en diversas tareas, desde la automatización hasta la gestión de capital humano y la supervisión del rendimiento, los procedimientos de contratación, el empleo de *cobots* y *chatbots* o el seguimiento de actividad (Moore, 2023).

Dada la amplísima variedad de funcionalidades inherentes a estos sistemas, resulta enormemente difícil concretar el impacto específico de cada uno de ellos, siendo evidente la necesidad de analizar cada caso específico por separado, considerando no solo sus ventajas, sino también sus respectivos inconvenientes, limitaciones y riesgos.

En los últimos años, los modelos de transformador (Vaswani *et al.*, 2017) han logrado resultados de vanguardia en una amplia variedad de tareas relacionadas con el lenguaje natural (Dai *et al.*, 2019; Radford *et al.*, 2019) y la comprensión discriminativa del lenguaje (Devlin *et al.*, 2019). Ahora bien, aunque el sector jurídico constituye uno de los sectores tradicionalmente más dominados por el uso del lenguaje, el lenguaje jurídico presenta características únicas, diferentes a las empleadas por el lenguaje coloquial, como términos poco comunes o inusuales, frases largas, expresiones antiguas o con significados propios a esta disciplina o que van aparejados a ciertas previsiones jurídicas (Chalkidis, Fergadiotis, Malakasiotis *et al.*, 2020; Zhong, Xiao *et al.*, 2020), hasta el punto de que diversos autores han llegado a considerarlo un sublenguaje (Tiersma, 1999; Williams, 2005). A ello se añade la elevada extensión de los textos jurídicos, más amplios que los habitualmente empleados para el entrenamiento de la mayoría de los modelos de aprendizaje profundo (Beltagy *et al.*, 2020; Chalkidis *et al.*, 2022; Hegel *et al.*, 2021; Zaheer *et al.*, 2021), existiendo ciertas dificultades en la clasificación y el etiquetado (Chalkidis, Fergadiotis, Malakasiotis *et al.*, 2020; Galgani *et al.*, 2012; Lippi *et al.*, 2019; Mencia y Furnkranzand, 2010; Nallapati y Manning, 2008; Tuggener *et al.*, 2020), lo que plantea importantes retos no solo en cuanto a la creación de *datasets* especializados, sino también para la capacitación y el ajuste de los modelos.

Desde el punto de vista técnico, numerosas investigaciones señalan el elevado potencial de los sistemas de inteligencia artificial basados en el lenguaje natural para el desarrollo de algunas actividades que podrían calificarse como de contenido propiamente jurídico (Aletras *et al.*, 2019, 2020; Ambrogi , 2023; Bommarito *et al.*, 2018; Chalkidis y Kampas, 2019; Chalkidis, Fergadiotis, Malakasiotis *et al.*, 2020; Chalkidis *et al.*, 2022; Kalson, 2022;

Perlman, 2023; Zhong, Xiao *et al.* 2020). Centrándonos en aquellas funcionalidades que podrían resultar de utilidad en la práctica jurídica, se incluyen, entre una amplísima variedad de aplicaciones, y de forma no exhaustiva, aquellas relacionadas con la búsqueda y la síntesis de jurisprudencia (Bhattacharya *et al.*, 2019, 2021; Jackson *et al.*, 2003; Locke y Zuccon, 2022; Tran *et al.*, 2019); la extracción de información de contratos (Chalkidis *et al.*, 2017, 2018, 2019; Hendrycks *et al.* 2021), incluyendo la identificación de riesgos respecto a su clausulado (Chalkidis *et al.*, 2017; Chen *et al.*, 2020; Gao *et al.*, 2012; Hendrycks, D. *et al.*, 2021; Leivaditi *et al.*, 2020; Lippi *et al.*, 2019); anonimización de documentos, permitiendo ocultar o reemplazar caracteres en documentos para reducir riesgos de localizar a personas o utilizar sus datos de manera fraudulenta; la minería de texto, para identificar nuevas tecnologías y medir la novedad de las patentes en el momento de su presentación, así como el impacto de estas nuevas tecnologías en la innovación posterior (Arts *et al.*, 2021); la preparación de juicios, vistas, interrogatorios y otras fases de prueba (Zhong, Wang *et al.*, 2020), incluyendo las actividades de descubrimiento o *e-discovery*; y la redacción de contratos, acuerdos u otros escritos. A las anteriores, se añaden las actividades de resolución de consultas jurídicas específicas (Kien *et al.* 2020; Navarro, 2023) –así en materia de políticas de privacidad (Ravichander *et al.*, 2019)–, la redacción de borradores o contratos, u otras funciones, como la traducción de documentos, la automatización de tareas repetitivas, la creación de *chatbots* basados en esta tecnología o, incluso, el desarrollo de *smart contracts*.

Un importante número de estudios se centra en el ajuste para la predicción de resultados de un procedimiento judicial o administrativo a través de diferentes fórmulas (Ferro *et al.* 2019; Medvedeva *et al.*, 2018, 2020). Entre las principales líneas de investigación, se encuentran aquellas centradas en la identificación de violaciones de derechos humanos (Aletras *et al.*, 2016; Chalkidis *et al.*, 2019); la predicción, en procedimientos penales chinos, de cargos penales, artículos aplicables y la duración de la pena (Hu *et al.* , 2017; Long *et al.*, 2019; Luo *et al.*, 2017; Xiao *et al.*, 2018; Yang *et al.*, 2019; Zhong *et al.*, 2018); o la predicción de resultados de casos ante el Tribunal Supremo de Alemania (Urchs *et al.*, 2021), Estados Unidos (Katz *et al.*, 2017; Kaufman *et al.*, 2019; Ruger *et al.*, 2004), Francia (Şulea *et al.*, 2017), Filipinas (Virtucion *et al.*, 2018), Reino Unido (Strickson y De la Iglesia, 2020), Suiza (Niklaus *et al.*, 2021), Turquía (Mumcuoğlu *et al.*, 2021), Tailandia (Kowsrihawat *et al.*, 2018), entre otros, existiendo, igualmente, investigaciones destinadas a interpretar las decisiones judiciales (Branting *et al.*, 2021; Chalkidis *et al.*, 2021; Ye *et al.*, 2018).

No obstante, conviene tener presente que un importante número de funcionalidades anteriores se encuentran todavía sujetas a importantes limitaciones técnicas, además de estar condicionadas a los pertinentes ajustes. En este sentido, algunos autores señalan que ChatGPT y otros sistemas de inteligencia artificial generativa tendrán, al menos a corto o medio plazo, un efecto limitado en el modelo de servicios legales (Adams, 2022; Bacas, 2022; Sellick, 2022). La propia OpenAI señala en sus políticas de uso de 10 de enero de 2024 que los usuarios no deben emplear los modelos para brindar asesoramiento jurídico sin que un profesional con la debida cualificación revise la información. Aun en tal caso, también señala la necesidad de advertir del uso de la asistencia de inteligencia artificial y de sus

posibles limitaciones. Asimismo, señala explícitamente que los modelos no están ajustados para ofrecer asesoramiento jurídico y que el usuario no debe confiar en los mismos como única fuente de asesoramiento legal OpenAI (2023a, 2024b).

Aunque resulte difícil anticipar de qué modo las soluciones basadas en modelos como GPT-4 y otros sistemas de inteligencia artificial generativa afectarán a la profesión jurídica y, particularmente, al sector de la abogacía, diversos informes, como el realizado por Accenture (2021) respecto a la distribución del tiempo de trabajo en el sector legal y el posible impacto de la inteligencia artificial, señalan que el 33 % de las actividades cuentan con un alto potencial de automatización, mientras que otros, como el de Goldman Sach, estiman que el 44 % de las tareas podrán ser automatizadas con inteligencia artificial (Hatzius *et al.*, 2023).

En la práctica, firmas como Allen & Overy (2023) o PwC (2023) han comenzado a implementar soluciones como Harvey –el sistema de inteligencia artificial generativa basado en GPT-4 y ajustado con datos específicos del sector legal– para el desarrollo de diversas funciones, a los que se sumaría recientemente Cuatrecasas con su modelo personalizado basado en Harvey, denominado Cuatrecasas Expert Legal IA (CelIA) (Cuatrecasas, 2023; Expert.AI, 2023; Sánchez Aristi *et al.*, 2023), lo que advierte de un posible cambio de paradigma en nuestra concepción de la profesión jurídica.

4. Limitaciones y riesgos en cuanto a su empleo en el sector jurídico

La implementación de sistemas de inteligencia artificial en el sector jurídico, incluyendo aquellas soluciones de inteligencia artificial generativa, como ChatGPT, o soluciones API basadas en modelos de la familia GPT, como GPT-4, obligará a tener en cuenta sus limitaciones y la posibilidad de alucinaciones y sesgos, así como sus riesgos, entre los que destacan aquellos aspectos relacionados con la protección de datos y la confidencialidad, la protección de derechos de propiedad intelectual e industrial y los derechos de autor. Además, habrá que tener en cuenta otros riesgos asociados, entre los que cabe citar, de forma no exhaustiva, aquellos derivados de la aplicación de la normativa en materia de publicidad o de competencia desleal o las posibles implicaciones éticas y deontológicas de su implementación.

4.1. Alucinaciones y sesgos

A pesar de los avances en el campo de los LLM, los diferentes modelos inspirados en esta tecnología, incluidos aquellos más avanzados, como GPT-4 (Bubeck *et al.*, 2023), continúan experimentando importantes limitaciones, entre las que se incluye la existencia de sesgos o la manifestación de alucinaciones, siendo tales aspectos y la falta de fiabilidad

algunas de las principales preocupaciones a abordar por parte las diferentes iniciativas regulatorias en curso (Cath *et al.*, 2018), así como por los principales estándares en el sector (IEEE, 2017; OCDE, 2019).

En el contexto de modelos, los «sesgos» se definen como la presencia de inexactitudes sistémicas, errores de atribución o distorsiones de hechos que dan lugar a que se favorezca a determinados grupos o ideas, a que se perpetúen estereotipos o a que se formulen suposiciones incorrectas basadas en patrones aprendidos (Ferrara, 2023). Entre los principales tipos y causas, distinguimos sesgos demográficos, raciales, culturales, de género, de edad, religiosos o socioeconómicos –cuando en los datos de entrenamiento existe sobrerrepresentación o subrepresentación de ciertos grupos demográficos, conduciendo al modelo a exhibir comportamientos sesgados o a perpetuar o reforzar tales estereotipos y prejuicios hacia los mismos–; sesgos lingüísticos –cuando los datos de entrenamiento se presentan mayoritariamente en un idioma, de forma que el modelo presenta un rendimiento sustancialmente diferenciado en detrimento de otros idiomas–; sesgos temporales –cuando los modelos, al estar restringidos a un periodo de tiempo determinado, conducen al mismo a una limitada comprensión de contextos históricos, a la sobrerrepresentación de ciertos hechos o a la expresión de información obsoleta–, así como sesgos de confirmación, ideológicos o políticos –cuando, a partir de los datos de entrenamiento, los modelos proporcionan salidas que refuerzan ciertas perspectivas específicas– (Ferrara, 2023). Un ejemplo de sesgo en modelos dirigidos a asistir en tareas jurídicas podría darse cuando el modelo asociara inadvertidamente ciertos aspectos legales con particulares factores demográficos, perpetuando estereotipos y potencialmente influyendo en las decisiones de los profesionales del derecho que emplearan estos sistemas. A reforzar y propagar los referidos sesgos contribuyen, además, ciertas aptitudes, como la generalización e inferencia, la aparición de capacidades emergentes o las propias actividades de corrección de los modelos por revisores humanos, las cuales también deberán ser supervisadas, pues podrían dar lugar a la aparición de nuevos sesgos.

Respecto a la detección y mitigación de sesgos, entre las principales medidas técnicas adoptadas se encuentra la curación y anotación de datos de entrenamiento, a fin de contar con datos de alta calidad y diversidad que permitan identificar y corregir algunos de los sesgos más importantes y la reducción de su influencia. Asimismo, el adecuado control y guiado durante las fases SFT y RLHF, en conjunción con una mejora en los mecanismos de evaluación y corrección de los modelos y del empleo de sistemas de recompensa basados en reglas, contribuyen a esta reducción y mitigación de sesgos. Sin embargo, aunque el empleo de datos de alta calidad y de los enfoques de intervención humana puedan ayudar a esta mitigación, resulta esencial reconocer las limitaciones existentes, incluyendo, especialmente, las relativas a la imposibilidad de eliminar la existencia de sesgos por completo. Uno de los aspectos más problemáticos a este respecto se centra, curiosamente, en la existencia de los propios sesgos humanos en la interpretación de los resultados, cuando los usuarios aceptan, bien por un exceso de confianza en los modelos o por falta de la adecuada supervisión, los resultados ofrecidos por los mismos.

La elevada coherencia de GPT-4 y la posibilidad de generar contenidos de forma más convincente y creíble que los modelos GPT anteriores –por ejemplo, debido al tono autoritario o a la presentación de información de forma aparentemente detallada– refuerzan su capacidad persuasiva y la atribución a los mismos de un «principio de autoridad», lo que aumenta significativamente el riesgo de exceso de confianza en que los resultados ofrecidos por los mismos son ciertos e inamovibles (OpenAI, 2023c). Es por ello que los proveedores que implementen este tipo de tecnologías en sus procesos productivos deberán garantizar la adecuada capacitación de su personal sobre la existencia de sesgos, su identificación y las limitaciones de los modelos, estableciendo un régimen adecuado de políticas de uso, cumplimiento y supervisión humana de los resultados, a fin de proporcionar retroalimentación al sistema y evitar el empleo de salidas sesgadas que pudieran derivar en posibles responsabilidades frente a terceros.

Deberá garantizarse que los datos utilizados para el desarrollo, la capacitación y el ajuste de los modelos cuenten con suficiente representatividad y diversidad, abordando el empleo de estos sistemas con un adecuado grado de transparencia sobre metodologías, fuentes de datos y limitaciones del modelo con el objetivo de permitir que los usuarios cuenten con una adecuada comprensión de los factores que puedan llegar a influir en las predicciones y decisiones de los mismos. Asimismo, resultará conveniente el establecimiento de marcos contractuales adecuados que determinen mecanismos de control, auditoría y evaluación de los modelos (Raji *et al.*, 2020), incluyendo, por ejemplo, la evaluación del rendimiento a partir de métricas de equidad y fiabilidad, o el empleo de técnicas de eliminación de sesgo algorítmico (Bender y Friedman, 2018; Dev y Phillips, 2019; Zhang *et al.*, 2018).

Por otra parte, los modelos pueden manifestar «alucinaciones», definidas como la generación de textos o respuestas que, a pesar de mantener un cierto grado de corrección gramatical, fluidez o apariencia de autenticidad, se desvían o entran en conflicto con el contenido de la fuente (alucinaciones intrínsecas, de factualidad o veracidad) (Lin *et al.*, 2022; Maynez *et al.* 2020), aunque se apoyara en otros supuestos (George y Stuhlmüller, 2023), o bien cuando, ante la falta de alineación factual, no pueden ser verificados a partir del contenido de una fuente (alucinaciones extrínsecas o de fidelidad) (Bruno *et al.* 2023; Guan *et al.*, 2023; Huang *et al.*, 2023; Ji *et al.*, 2023). Las primeras se proyectan, entre otros, a través de inconsistencias factuales, cuando el modelo expresa hechos que, aun pudiendo fundamentarse en información real, presentan contradicciones o errores; o bien a través de la creación o fabricación factual, cuando este expresa información no verificable frente a un conocimiento real establecido (Huang *et al.*, 2023). Por su parte, las alucinaciones de fidelidad tienen lugar cuando el modelo exhibe contradicciones o inconsistencias, desviándose respecto a las instrucciones del usuario o del contexto proporcionado por el mismo (Adlakha *et al.*, 2023; Cao *et al.*, 2017; Liu, Iter *et al.*, 2023; Min *et al.*, 2023); o bien en su razonamiento interno, exhibiendo contradicciones lógicas, observadas con frecuencia en tareas de razonamiento o explicabilidad. Así, por ejemplo, al solicitar al modelo la cita de precedentes relacionados con un caso concreto o la elaboración del resumen de una

sentencia, las alucinaciones factuales podrían proyectarse, por ejemplo, en la distorsión de la información presente en su contenido, la fabricación de hechos o la manifestación de errores en la referencia de los precedentes, su completa fabricación o la expresión de información no atribuible a los mismos; mientras que las alucinaciones de fidelidad podrían traducirse en la desviación respecto a las instrucciones proporcionadas por el usuario.

Las alucinaciones responden a diferentes causas, sistematizadas conforme a diferentes criterios (Huang *et al.*, 2023). Aquellas que están relacionadas con datos derivan principalmente del empleo de fuentes de datos defectuosas –por incluir sesgos, errores o desinformación (Lin *et al.*, 2022), lo que puede conducir a falsedades imitativas que amplifiquen inadvertidamente inexactitudes o resultados engañosos, o bien a posibles sesgos de duplicación (Lee *et al.*, 2022) u otros sesgos sociales– y conocimientos limitados (Onoe *et al.*, 2022), incluyendo la utilización deficiente de tales datos, como ocurre en ocasiones ante el empleo de atajos (Kandpal *et al.*, 2023; Kang y Choi, 2023; Li *et al.*, 2022) y fallos en la recuperación de conocimientos (Zheng, 2023; Liu, Lin *et al.*, 2023; Mallen *et al.*, 2023).

Las limitaciones de conocimiento se hacen evidentes cuando los modelos se enfrentan a problemas en dominios especializados, como preguntas médicas (Li *et al.*, 2023; Singhal *et al.*, 2023) o jurídicas (Katz *et al.*, 2023; Yu *et al.*, 2022), resultando frecuente que, en tales casos, los modelos manifiesten alucinaciones pronunciadas, a menudo bajo la forma de creaciones factuales o la expresión de información desactualizada.

Por otra parte, durante las etapas de preentrenamiento, la existencia de defectos en la arquitectura del modelo como una representación unidireccional inadecuada –que dificulte su capacidad para captar dependencias contextuales– o las limitaciones en los módulos de atención, a efectos del razonamiento algorítmico, aumentarán el riesgo de alucinaciones.

Asimismo, se ha constatado que los sesgos de exposición durante el entrenamiento contribuyen a la aparición de alucinaciones y errores en cascada (Zhang, Press *et al.*, 2023). Aunque el alineamiento es crucial para mejorar las capacidades de los LLM, también introduce el riesgo de alucinaciones en los casos de desajuste de capacidades o creencias (Huang *et al.*, 2023), así como por las deficiencias en las estrategias de decodificación a causa de la excesiva aleatoriedad en el muestreo o la insuficiente atención al contexto (Chen *et al.*, 2022), incluyendo la excesiva dependencia del contenido de proximidad y que podrán limitar las probabilidades de salida (McKenna *et al.*, 2023).

Aunque GPT-4 cuente con mecanismos más desarrollados para prevenir la aparición de alucinaciones que sus predecesores (OpenAI, 2023c), la posibilidad de su manifestación continúa siendo una de las principales preocupaciones respecto a su fiabilidad (Huang *et al.*, 2023), particularmente en sectores especializados como el jurídico, en los que la fiabilidad resulta fundamental. Más aún, la posibilidad de que tales errores y alucinaciones puedan entremezclarse con información en apariencia correcta contribuye a dificultar su identificación (Bubeck *et al.*, 2023).

Como ya adelantamos, la mayor capacidad de persuasión y coherencia de modelos como GPT-4 frente a sus predecesores (OpenAI, 2023c) conduce al agravamiento del precitado sesgo por exceso de confianza, lo que redundará en una menor comprobación o verificación, por parte de los usuarios, de la veracidad de las respuestas proporcionadas por los modelos y, en consecuencia, en un mayor grado de indetección de alucinaciones. Esta potencial confianza acrítica dará lugar a situaciones de espejismo cognitivo (*cognitive mirage*), contribuyendo a la adopción de decisiones equivocadas y a una cascada de consecuencias no deseadas (Ye *et al.*, 2023; Zhang, Li *et al.*, 2023) y que, desde el punto de vista de su implementación en el campo jurídico, podrán conducir, en última instancia, a supuestos de responsabilidad frente a terceros, así como a posibles responsabilidades deontológicas. En Estados Unidos, en el asunto Mata versus Avianca, Inc., varios abogados de la firma Levidow & Oberman fueron sancionados por mala fe y negligencia tras el empleo de jurisprudencia falsa creada por ChatGPT (CBS News, 2023; Merken, 2023). En otro asunto, Brantley Starr, el juez del distrito norte de Texas ordenó que los abogados se comprometieran a no utilizar ChatGPT u otra tecnología de inteligencia artificial generativa para escribir informes legales y ello ante la posibilidad de que la citada tecnología pudiera inventar hechos (Cerullo, 2023b); fundamentación que se apoyaría en otro asunto anterior, por el que un abogado preparó un informe que presentó ante el juez, el cual estaba basado en una investigación realizada por ChatGPT (Cerullo, 2023a).

A nivel técnico, existen diferentes estudios relacionados con la detección y mitigación de los efectos de las alucinaciones. Entre las principales estrategias de detección, se encuentran aquellas relativas a la recuperación de datos externos y a la recopilación de evidencias, comparando el contenido generado por el modelo con fuentes de conocimiento confiables y que, sin perjuicio de sus limitaciones (Guo *et al.*, 2022), actuarán como sistemas de verificación, fomentando un mayor grado de explicabilidad de los resultados. Otras estrategias se orientan, en cambio, a la estimación del grado de incertidumbre del contenido factual generado por el modelo a partir de los estados internos del LLM, con métricas como la probabilidad mínima de *tokens* o la entropía, o bien mediante sistemas de autoevaluación. La detección de alucinaciones de fidelidad se enfoca en garantizar el correcto alineamiento del contenido generado a partir del contexto dado, a fin de evitar, por ejemplo, respuestas irrelevantes o contradictorias, empleando métricas basadas en hechos, clasificadores, sistemas de preguntas y respuestas o de estimación de incertidumbre, así como enfoques basados en indicaciones. Por otra parte, frente a las alucinaciones relacionadas con los datos de entrenamiento, destacan las estrategias para abordar sesgos, información errónea y lagunas, incluyendo la importancia de mantener la corrección fáctica de los datos y de eliminar duplicaciones y sesgos sociales. Frente a las alucinaciones derivadas de limitaciones de conocimiento, además de un adecuado control sobre el procedimiento de ajuste, destaca la exploración de enfoques –como la edición de conocimiento, que busca corregir el comportamiento del modelo mediante la incorporación de información adicional– y la generación aumentada por recuperación (Savelka *et al.*, 2023). Frente a las alucinaciones derivadas del entrenamiento de los modelos, se exploran soluciones dirigidas a abordar las limitaciones en su arquitectura y los errores durante la fase de atención, la mejora en las

estrategias de entrenamiento y la limitación del sesgo de exposición. Finalmente, respecto a la mitigación de alucinaciones relacionadas con el desalineamiento, se exploran estrategias como el muestreo del núcleo factual, intervención en tiempo de inferencia, así como estrategias avanzadas para refinar el proceso de decodificación y mejorar la factibilidad y fidelidad de los resultados generados.

No obstante, y sin perjuicio de los importantes avances técnicos en la detección y mitigación de alucinaciones, de sesgos y de otros errores, el uso responsable y cauto de modelos como ChatGPT u otros basados en la familia GPT, consistente en contrastar y verificar los resultados ofrecidos por los modelos, especialmente en el sector jurídico, continúa siendo un principio general que ha de inspirar las políticas internas que se estipulen sobre su uso.

4.2. Riesgos en materia de protección de datos

El desarrollo e implementación de sistemas de inteligencia artificial como ChatGPT, así como de aquellas soluciones basadas en modelos como GPT-4, plantea importantes retos y cuestiones en materia de cumplimiento normativo y, particularmente, en materia de protección de datos. El análisis y la evaluación de los riesgos asociados conforme a la normativa de protección de datos europea, destacando el Reglamento 2016/679, de 27 de abril de 2016, relativo a la protección de las personas físicas en lo que respecta al tratamiento de datos personales y a la libre circulación de estos datos (RGPD), constituye una tarea de enorme complejidad, dada no solo la pluralidad de fases, particularidades y riesgos inherentes a cada una de las etapas que integran el ciclo de vida de un sistema de inteligencia artificial determinado (De Silva y Alahakoon, 2021) –durante las fases de entrenamiento, validación, despliegue, explotación, inferencia, decisión y ajuste del modelo, así como posibles transmisiones a terceros, en función de las características y de la solución particular ante la que nos encontremos (Agencia Española de Protección de Datos [AEPD], 2020)–, sino también las características propias de cada sistema o solución de inteligencia artificial específica, lo que dificulta sustancialmente su concreción, obligando a adoptar un enfoque holístico, responsable y orientado al riesgo y a tener presente, de forma permanente, si las diferentes actividades ante las que nos encontremos conllevan o no una actividad de tratamiento de datos personales o si, en su caso, forman parte de un tratamiento más amplio.

A tal fin, deberán tenerse en cuenta una amplia variedad de aspectos relacionados con el diseño, la configuración, la capacitación, el funcionamiento y la explotación del modelo; el contexto y la específica función que el mismo vaya a desempeñar –debiendo tener en cuenta, por ejemplo, si el mismo será o no empleado en la toma de decisiones automatizadas o en la elaboración de perfiles–, a los que habrán de añadirse otros elementos, como las particularidades inherentes a la interacción de los usuarios con el sistema o las políticas de privacidad y condiciones contractuales y de servicio de sus proveedores, entre otros múltiples aspectos.

Asimismo, respecto a las actividades de tratamiento, deberá prestarse especial atención, entre otros, a los aspectos relacionados con la legitimación para dicho tratamiento, a los sujetos intervinientes y a los flujos de datos, a las particulares obligaciones en materia de información y transparencia, a las garantías en cuanto al ejercicio de derechos y seguridad, a las transferencias internacionales de datos, a la evaluación de impacto y al análisis de la proporcionalidad del tratamiento, incluyendo, cada vez más, el análisis del marco de tratamiento a partir no solo de la normativa en materia de protección de datos y de los derechos y valores sociales consagrados en la carta de la Unión Europea, en los tratados de la Unión Europea y en las constituciones nacionales, sino también de principios éticos, como los de autonomía, prevención de daños, equidad y explicabilidad (AEPD, 2020; Sartor, 2020).

Las actividades de preentrenamiento de los modelos de la familia GPT, incluido GPT-4, se efectúan en gran medida a partir del *dataset* conformado por vastos corpus de texto escrito procedente de diferentes fuentes (OpenAI, 2023c), como Common Crawl, WebText2, Books1 y Books2 o Wikipedia. La propia actividad de *web scraping* de los cientos de miles de webs de las que se extrae la información que integra los referidos corpus, cuando las mismas contengan datos personales –así como, en su caso, las posibles actividades de preprocesamiento de información, tratamiento de datos no estructurados, limpieza, balanceo, selección, transformación, partición del conjunto de datos para verificación e información de trazabilidad o de auditoría–, se calificarán como actividades de tratamiento, a efectos de la normativa en materia de protección de datos, lo que exigirá que las mismas reúnan los requisitos de legitimación previstos por la normativa, así como el respeto a los principios relativos al tratamiento, estableciendo las salvaguardias técnicas y organizativas necesarias para proteger y gestionar dichos datos personales.

No obstante, aunque los datos personales hubieran sido obtenidos de una fuente públicamente accesible, ello no serviría, por sí mismo, como base jurídica para el tratamiento, no solo ante la inexistencia de un concepto legal de «fuentes accesibles al público» (AEPD, 2018, 2021a, 2021b, 2023a), sino por la prevalencia del consentimiento específico como base jurídica del tratamiento, de forma que dicha accesibilidad pública operaría simplemente como un elemento más en la ponderación del interés legítimo, siendo necesario que concurra alguna de las causas legitimadoras del artículo 6 del RGPD.

OpenAI reconoce la inclusión de información personal entre los datos de capacitación de sus modelos a partir de los referidos corpus (OpenAI, 2023c, 2024c), señalando que los datos personales incluidos en las interacciones de los usuarios durante la utilización de los servicios –y, por tanto, en el contenido de las entradas o *prompts*, cargas de archivos o comentarios–, así como la información de cuentas, comunicaciones y redes sociales, eventos y encuestas; información técnica de registro, uso, dispositivo y *cookies* cuando el usuario visita la página web y utiliza o interactúa con los servicios (OpenAI, 2023c, 2023d, 2023e), incluyendo la información disponible públicamente en internet o proporcionada por terceros, podrá ser empleada para proporcionar, mantener, mejorar y desarrollar los servicios de OpenAI, entre los que se integran los modelos GPT, como GPT-4, y las aplicaciones como ChatGPT.

Asimismo, tales datos podrán ser empleados por OpenAI para comunicarse con el usuario y para cumplir con las obligaciones legales y proteger los derechos, la privacidad, la seguridad o la propiedad de los usuarios, de sus afiliados o de cualquier tercero.

A pesar de que, desde el 15 de febrero de 2024, el tratamiento de datos personales de usuarios residentes en el Espacio Económico Europeo y Suiza sea llevado a cabo por la filial irlandesa OpenAI Ireland Limited, la propia OpenAI advierte de la transferencia internacional de datos personales fuera del Espacio Económico Europeo para los fines descritos en su política de privacidad[2], basándose en las decisiones de adecuación de la Comisión Europea sobre determinados países y, en el caso de otras jurisdicciones, en las Cláusulas Contractuales Tipo aprobadas por la Comisión Europea y en las Adendas aplicables a cada país (OpenAI, 2023e). Todo ello sin perjuicio de las previsiones contenidas en materia de privacidad empresarial, circunscrita exclusivamente a la utilización de sus modelos comerciales –ChatGPT Team, ChatGPT Enterprise y la plataforma API–, por las que OpenAI señala que los datos introducidos durante la utilización de estos modelos no serán empleados con finalidades de capacitación, añadiendo que los usuarios de la versión ChatGPT Enterprise podrán controlar la duración de conservación de dichos datos. A este respecto, OpenAI señala que cualquier conversación eliminada durante la utilización de ChatGPT Enterprise será eliminada de sus sistemas dentro de los 30 días siguientes, a menos que estuviera obligada legalmente a conservarla. La misma previsión se estipula respecto al uso de servicios API, sin perjuicio de que esté permitido el acceso a dicha información para fines de investigación de posibles abusos de la plataforma y cumplimiento legal, incluyendo a terceros contratistas externos especializados, sujetos a obligaciones de confidencialidad y seguridad, para dicha finalidad (OpenAI, 2024a).

Centrándonos en el tratamiento de datos personales, entre las bases jurídicas de tratamiento especificadas por OpenAI se incluyen las relativas al consentimiento, la ejecución de la relación contractual, la protección de los intereses legítimos o el cumplimiento de las obligaciones legales, entre otros, en función de la actividad específica. Ahora bien, incluso admitiéndose la existencia de una base jurídica, ello no habilitará al uso de los datos para cualquier propósito y en todo momento, sino que deberá restringirse a fines determinados, legítimos e identificados, evitando su tratamiento de manera incompatible.

[2] Respecto a las transferencias internacionales, OpenAI señala que, en cumplimiento de la política de privacidad, dichas transferencias se llevarán a cabo basándose en las decisiones de adecuación de la Comisión Europea en ciertos países y, para otras jurisdicciones, en las cláusulas contractuales estándar aprobadas por la Comisión Europea, en línea con el artículo 46.2 del RGPD, como las incluidas en la Decisión de Ejecución (UE) 2021/914 de la Comisión de 4 de junio de 2021. Teniendo en cuenta que los datos personales se procesarán y almacenarán en las instalaciones y servidores de OpenAI en Estados Unidos, la adopción de cláusulas contractuales tipo para la transferencia de datos personales a terceros países constituye una medida apropiada siempre que OpenAI pueda cumplir con su contenido, y ello ante la ausencia de garantías equivalentes a las europeas en la regulación estadounidense, como se concluyó en Sentencia del Tribunal de Justicia (Gran Sala) de 16 de julio de 2020, Comisaria de Protección de Datos versus Facebook Irlanda y Maximillian Schrems (Schrems II), C-311/18 (Jelinek, 2020).

Los interesados cuyos datos sean tratados deberán ser conscientes de cómo van a utilizarse, en consonancia con los principios de información y transparencia; debiendo depurarse, conforme al principio de minimización, toda la información no estrictamente necesaria para el entrenamiento de los modelos, eliminando el resto, a menos que se justifique la necesidad de mantenerlos para el refinado o evaluación del sistema, o se justifique la necesidad y legitimidad de mantenerlos para otras finalidades compatibles.

Más aún, la invocación del interés legítimo como base jurídica, reclamará del responsable un mayor grado de compromiso, formalidad y competencia, así como una cuidadosa evaluación de que los intereses legítimos prevalecen sobre el posible impacto de los derechos, las libertades y los intereses de los interesados. Respecto a tales casos, además, deberán ponderarse las eventuales medidas compensatorias derivadas de mantener el tratamiento bajo supervisión continua; el grado de responsabilidad proactiva; la incorporación de medidas de privacidad o la aplicación de buenas prácticas, como la de posibilitar la opción de *opt-out* a los interesados (AEPD, 2020). El responsable deberá demostrar a las autoridades que el impacto de este tratamiento no es tan significativo como para impedir el mismo sobre esa base, debiendo quedar documentado todo este proceso de análisis y toma de decisión en cumplimiento del principio de responsabilidad (*accountability*). Con todo, aun cuando el tratamiento se fundamente en el interés legítimo –y, por tanto, no sea necesario recabar el consentimiento del interesado–, las obligaciones de información previstas en los artículos 13 y 14 del RGPD continuarán aplicándose.

A los riesgos anteriores se suman aquellos otros derivados de las actividades de despliegue –cuando la solución de inteligencia artificial incluya el acceso a datos personales o exista forma de obtenerlos, por ejemplo, a través de patrones o dentro de la lógica del modelo–; explotación e inferencia –cuando se usen datos del interesado o de terceros para obtener un resultado, o bien cuando los datos y las inferencias del interesado se almacenan–; y comunicación, almacenamiento o acceso a terceros (AEPD, 2020), lo que advierte del elevado riesgo de incumplimiento de la normativa de protección de datos.

OpenAI ha manifestado su compromiso para reducir el procesamiento de información personal durante la fase de entrenamiento de sus modelos mediante la agregación o desidentificación de datos personales a efectos de su seudonimización y anonimizacion; la eliminación de datos personales y aquellos procedentes de sitios web que agregan grandes volúmenes de información personal; la implementación de medidas técnicas, administrativas y organizativas para proteger la información personal contra la pérdida, uso indebido, acceso, divulgación, alteración o destrucción no autorizados, así como la realización de autoevaluaciones o el entrenamiento de los modelos para rechazar solicitudes de información privada o sensible sobre personas (OpenAI 2023d, 2023e). Sin embargo, lo cierto es que todavía se encuentran pendientes un importante número de investigaciones por parte de diferentes autoridades gubernamentales en materia de protección de datos respecto a la suficiencia de tales medidas.

Asimismo, deberá tenerse en cuenta la previsión contractual de divulgación de datos personales por parte de OpenAI a vendedores y proveedores de servicios o como parte de procedimientos de *due diligence* en el seno de transferencias u otras transacciones estratégicas, incluyendo sus obligaciones legales de informar a las autoridades gubernamentales u otros terceros y afiliados, así como, en su caso, a los administradores de cuentas comerciales. Aunque OpenAI haya puesto a disposición de los usuarios un formulario para la oposición al procesamiento de datos personales y su eliminación de los datos de salida de ChatGPT (Markovski, 2023; OpenAI, 2024c, 2024d, 2024e), esta reconoce que, de conformidad con la normativa de protección de datos, algunos derechos podrían no ser absolutos, no garantizando que la información se elimine de los resultados de ChatGPT. Sin embargo, diversos estudios señalan que el entrenamiento y la capacitación de un modelo a partir de información que incluya datos personales podría permitir al mismo acceder a dicha información, y ello a pesar de que los datos hubieran sido anonimizados o depurados (Lomas, 2019), incluso mediante el empleo de mecanismos de inhibición de comportamientos específicos, filtros a nivel de salida o medidas RLHF para rechazar solicitudes que incluyan datos personales (Meng *et al.*, 2023; Patil *et al.*, 2023). Estas conclusiones plantean importantes preocupaciones no solo respecto a la posibilidad de acceso a dicha información por parte de terceros y otros usuarios de modelos como GPT-4, sino también en cuanto a la posibilidad de garantizar al interesado o titular de tales datos personales la efectividad de sus derechos de supresión y rectificación, especialmente cuando tales datos se encuentren incrustados en el modelo.

Centrándonos en el uso de ChatGPT Enterprise o de soluciones API basadas en modelos como GPT-4 –entre los que habrán de incluirse, por ejemplo, aquellos supuestos de implementación comercial de los modelos en la práctica jurídica, como podría ser la relativa a las actividades de asesoramiento por parte de un despacho de abogados–, deberá determinarse, en primer lugar, si el uso que vaya a darse a estos sistemas conllevará el tratamiento de datos personales, pues resulta evidente que no todas las soluciones basadas en inteligencia artificial implicarán una actividad de tratamiento ni, en su caso, esta alcanzará todas las fases de su ciclo de vida. En tales casos, serán precisamente dichas actividades de tratamiento las que se encuentren sujetas a la regulación en materia de protección de datos, mientras que, si una actuación concreta no implica una actividad de tratamiento de datos personales, no quedará sujeta a la misma.

Cuando la función que desempeñe el sistema de inteligencia artificial conlleve el tratamiento de datos personales, o bien cuando el sistema constituya un elemento del procesamiento de datos u otro tipo de operación dentro de una o más de las fases que integren una actividad más amplia de tratamiento –en cuyo caso, la función del sistema de inteligencia artificial no se considerará como un tratamiento aislado, sino como una operación dentro de dicha actividad de tratamiento más amplia, pues la funcionalidad aislada podría no legitimarse si no se incluye en un tratamiento amplio con una finalidad última y legítima (AEPD, 2023b)–, dicho tratamiento deberá cumplir con el conjunto de previsiones contenidas en la normativa en materia de protección de datos. Así, deberá inspirarse en los prin-

cipios de licitud, lealtad, transparencia, limitación en su finalidad, minimización, exactitud, limitación de plazo de conservación, integridad y confidencialidad (art. 5 del RGPD), bajo el prisma del principio de responsabilidad proactiva por parte del responsable del tratamiento. Asimismo, deberá contar con una base jurídica legitimadora (arts. 6 a 11 del RGPD), la cual podrá fundarse en la ejecución de un contrato en el que el interesado sea parte o en la aplicación de medidas precontractuales, en el interés legítimo, en el propio consentimiento de los interesados, en razones de interés público o en el cumplimiento de obligaciones legales, entre otros. De igual modo, deberá prestarse especial atención a la posibilidad de tratamiento de alguna de las categorías especiales de datos y a las posibilidades de enervación de la prohibición de su tratamiento a través de alguna de las excepciones previstas (art. 9.2 del RGPD). Será necesario velar por el cumplimiento de las obligaciones de información a los interesados (arts. 12 a 14 del RGPD), admitiéndose la información a través de un sistema de aproximación por capas o niveles. Así, deberá identificarse la identidad del responsable del tratamiento, la finalidad del mismo y la posibilidad del ejercicio de los derechos previstos en los artículos 15 a 22 del RGPD, incluyendo referencia explícita a si el tratamiento incluye la elaboración de perfiles o decisiones automatizadas –en cuyo caso, será preceptivo informar de esta circunstancia y de los derechos de oposición, proporcionando información significativa sobre la lógica aplicada y las consecuencias de dicho tratamiento–, añadiendo información sobre el carácter recuperable o no de los datos y, en caso de que no hubieran sido obtenidos directamente, sus categorías y fuentes, y dedicando una segunda capa a la información prevista en los artículos 13 y 14 del RGPD (AEPD, 2020).

Respecto al tratamiento en sí, un aspecto importante se centra en la correcta aplicación de los principios previstos por la normativa. Así, el principio de transparencia impone que toda información y comunicación relativa al tratamiento sea concisa, accesible y comprensible, a fin de garantizar un tratamiento leal y transparente y que los interesados tengan conocimiento de los riesgos, de las normas y de las salvaguardias del tratamiento. El principio de limitación determinará que la base jurídica del tratamiento deba restringirse a aquellos fines determinados, explícitos y legítimos que se hayan identificado, evitando tratarlos de forma incompatible con esos fines.

Por otra parte, se plantean ciertas cuestiones en cuanto a la forma de cohonestar las exigencias derivadas del principio de exactitud y calidad del dato con la posible existencia de alucinaciones, sesgos y errores en los modelos, ya que el mismo exige que los datos recogidos o inferidos a través de procedimientos matemáticos, estadísticos o para la elaboración de perfiles sean exactos –incluyendo la implementación de métricas y técnicas de depuración y trazabilidad para garantizar la fidelidad e integridad del conjunto de datos–, así como la obligación de documentar que los procedimientos empleados para el entrenamiento y la inferencia de la información sobre un interesado sean precisos, estables y predecibles. Conforme al principio de minimización, los datos personales solo deberán tratarse si la finalidad del tratamiento no pudiera lograrse razonablemente por otros medios, persiguiendo limitar la extensión de las categorías de datos a aquellas estrictamente necesarias, limitando la extensión en el número de interesados, así como también su accesibilidad al personal responsable o encargado y, particularmente, al usuario final y a terceros.

Asimismo, deberá garantizarse la posibilidad de ejercicio por los interesados de los derechos de acceso, rectificación, supresión, limitación, portabilidad y oposición, adoptando precauciones específicas en caso de su empleo en decisiones automatizadas o en la elaboración de perfiles (arts. 15 a 23 del RGPD). Deberán establecerse garantías y efectuarse evaluaciones de impacto cuando su naturaleza, alcance, contexto o fines entrañen un alto riesgo para los derechos y las libertades de las personas físicas (art. 35 del RGPD). Para ello, resultará preceptiva la llevanza de un registro de actividades de tratamiento, así como el cumplimiento de las obligaciones de documentación y gestión de riesgos frente a los derechos y libertades de los titulares de los referidos datos, incluyendo el establecimiento de medidas técnicas y organizativas que garanticen un nivel apropiado de seguridad (arts. 24 a 43 del RGPD). Del mismo modo, será preciso elaborar las evaluaciones de impacto oportunas y los avisos de privacidad, debiendo cumplir, en su caso, con las condiciones previstas para la realización de transferencias internacionales (arts. 44 a 50), así como con las previsiones contractuales específicas, como las relativas a la ejecución del anexo correspondiente con OpenAI sobre procesamiento de datos personales (OpenAI, 2024f). La implementación de sistemas de inteligencia artificial dentro de procesos productivos que conlleven el tratamiento de datos personales requerirá, en consecuencia, no solo del adecuado tratamiento, sino también de la elaboración y aplicación de los esquemas de gobernanza y supervisión oportunos, adaptados tanto al tipo de organización como al tipo de sistema de inteligencia artificial que vaya a implementarse y de su específica funcionalidad, debiendo contar con una adecuada regulación interna que establezca la forma en que los empleados y miembros de la organización traten los referidos datos e interactúen con los modelos, con el fin de evitar posibles infracciones en materia de protección de datos y supuestos de responsabilidad frente a terceros.

4.3. Derechos de propiedad intelectual e industrial y bases de datos

El auge de los sistemas de inteligencia artificial como ChatGPT ha suscitado importantes cuestiones a nivel global en torno a su interacción con la protección de los derechos de propiedad intelectual e industrial. A la complejidad inherente a este fenómeno, derivada, entre otros extremos, de la enorme variedad y diversidad de tecnologías que lo integran y de su rápida evolución, así como de las diferentes fases que integran el ciclo de vida de estos sistemas, se añaden aquellos aspectos relacionados con la propia naturaleza de los derechos de propiedad intelectual, de la globalización e internet, destacando, particularmente, las limitaciones territoriales del marco jurídico de protección y la falta de uniformidad entre ordenamientos jurídicos, sin perjuicio de la protección dispensada por ciertos instrumentos internacionales, como el Convenio de Berna para la Protección de Obras Literarias y Artísticas de 1886.

Entre las principales cuestiones en materia de propiedad intelectual se encuentran aquellas relacionadas con la elaboración de los corpus y datos de entrenamiento, así como su posterior utilización durante las fases de desarrollo y ajuste de los modelos. Ello resulta es-

pecialmente grave ante el riesgo de que tales corpus de texto contengan obras o fragmentos de obras protegidas que no se hallen en el dominio público y que podrían estar siendo utilizadas sin contar con autorización legal o sin el consentimiento o autorización de sus legítimos titulares, en infracción de los derechos morales y patrimoniales de los mismos. Del mismo modo, la generalización de estos sistemas ha dado lugar al surgimiento de cuestiones relacionadas con el funcionamiento técnico de los mismos, planteándose si nos encontramos ante actos de transformación, modificación o divulgación de obras protegidas cuando estas sean puestas a disposición de los usuarios, surgiendo también cuestiones respecto a la autoría tanto de los *prompts* de entrada como del *output* o resultado de salida y de las posibilidades de su explotación, incluyendo la incidencia del *prompt* en tales resultados.

El entrenamiento y la capacitación de los modelos, incluyendo sus fases de validación y ajuste con contenido protegido, sin contar con la suficiente legitimación para ello, plantea igualmente un riesgo evidente en materia de propiedad intelectual e industrial y de derechos de autor. Previamente, el empleo de rastreadores en línea (*web crawlers*) y el uso de actividades de *web scraping* en la conformación de los corpus de entrenamiento empleados en el preentrenamiento de los modelos de la familia GPT (OpenAI, 2023c, 2024c) plantea importantes riesgos en materia de propiedad intelectual y de derechos de autor. Una parte sustancial de la información de internet, incluida aquella disponible en abierto o sin restricción, se integra por contenido protegido por derechos de propiedad intelectual o industrial y derechos de autor, entre los que se incluyen obras completas o fragmentos de obras protegidas, composiciones de palabras, libros, artículos de periódicos y revistas, monografías, obras literarias, composiciones musicales, imágenes, películas, videojuegos, fragmentos de código o *software*, entre otros.

Más allá de las excepciones y limitaciones previstas en la Directiva 2001/29/CE –como la relativa a actos de reproducción provisional–, quienes hubieran llevado a cabo las actividades de *web scraping* solo podrían ampararse en una de las dos excepciones relativas a la minería de textos y datos previstas por la Directiva 2019/790, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital, objeto de transposición en España a través del Real Decreto-Ley 24/2021, de 2 de noviembre. La primera de ellas permite la reproducción y extracción a través de minería de textos y datos de obras a las que se tenga acceso lícito –incluyendo contenidos de acceso abierto o disponibles de forma gratuita en línea– cuando esta sea realizada por organismos con fines de investigación científica. Sin embargo, dada la organización corporativa de OpenAI, la existencia de fin de lucro, la incidencia en la misma por parte de Microsoft y la participación principal de esta última en los resultados de la investigación, parece cuestionable que la misma pueda ampararse en esta excepción. No obstante, la segunda excepción permite las reproducciones y extracciones de obras de forma legítima para fines de minería de textos y datos, al menos cuando no exista reserva expresa por los titulares de derechos, la cual podrá reflejarse a través de medios de lectura mecánica –como el estándar o protocolo de exclusión robots.txt– en el caso del contenido puesto a disposición del público en línea (García Vidal, 2020; Sánchez Aristi *et al.*, 2023).

Respecto al *web scraping* de bases de datos, conviene tener presente que las mismas podrán estar protegidas por derechos de autor en cuanto a su estructura cuando esta constituya una creación intelectual original de su autor. Aunque tal protección por propiedad intelectual no se hará extensiva a su contenido, se admite su protección por parte del autor de la base de datos a través del derecho *sui generis* previsto por la Directiva 96/9/CE, de 11 de marzo de 1996, así como, en su caso, por la legislación de trasposición. La extracción o la reutilización de la totalidad o de una parte sustantiva del contenido de una base de datos, cuando la obtención, la verificación o la presentación de dicho contenido hubiera representado una inversión sustancial desde el punto de vista cuantitativo o cualitativo, incluyendo las actividades de recuperación, rastreo, reutilización, búsqueda en tiempo real y copia local del contenido de las mismas, constituyen una potencial infracción en materia de bases de datos (Sentencia del Tribunal de Justicia de 15 de enero de 2015 [Ryanair Ltd contra PR Aviation BV, Case C30/14] y Sentencia del Tribunal de Justicia de 3 de junio de 2021 [SIA «CV-Online Letonia» contra SIA «Melons», C762/19]), cuando tales actividades no cuenten con suficiente legitimación o puedan incluirse en alguna de las excepciones previstas al derecho *sui generis*, a las que habrán de añadirse las excepciones previstas por la Directiva 2019/790, de 17 de abril de 2019, en los términos antes expuestos.

En cuanto a los riesgos asociados al *output* de salida de los modelos, algunas de las cuestiones planteadas conectan con los límites a los derechos de reproducción, la comunicación pública y la transformación de obras protegidas. En caso de que los modelos pongan a disposición de los usuarios respuestas que reproduzcan total o parcialmente el contenido de una obra protegida o que conlleve la traducción, adaptación o cualquier otra modificación de la misma, dando lugar a una obra diferente, estaremos ante un supuesto de reproducción o, en su caso, de transformación, el cual conllevará un elevado riesgo de infracción en materia de propiedad intelectual por parte de OpenAI.

Tales aspectos se encuentran en la base de un creciente número de demandas y reclamaciones contra OpenAI, fundamentalmente en Estados Unidos, como las formuladas recientemente por parte del *New York Times* o la demanda colectiva Author's Guild *et al.* versus OpenAi, Inc. *et al.* (1:23-cv-08292), en la que se incluyen diversos autores, como George R. R. Martin (Zahn, 2023); o la interpuesta por el autor Julian Sancton en el asunto Julian Sancton versus OpenAI, Inc. (1:23-cv-10211). No obstante, en Estados Unidos, todavía no existe jurisprudencia clara que describa la aplicación específica del uso legítimo (*fair use*), establecida por la sección 107 de la Copyright Act 1976 y reflejada en el título 17, capítulo 1, parágrafo 107, del United States Code en cuanto a los sistemas de inteligencia artificial generativa. En tales casos, es probable que el argumento de defensa que intente seguirse por parte de OpenAI sea similar al sostenido por Google en el asunto Authors Guild, Inc. versus Google, Inc. (13-4829-cv [2d Cir. 2015)].

De seguirse el criterio sostenido en este precedente, OpenAI podría intentar justificar que los fragmentos de texto protegido que proporciona queden, o bien amparados por el derecho de cita, o bien encuadrados dentro de un acto más amplio de transformación, amparado

por el *fair use* –según el concepto sostenido en el asunto Campbell versus Acuff-Rose Music, Inc. (510 US 569 [1994] 591)–, si bien para ello deberán acreditar que tales fragmentos ofrecen algo nuevo y diferente del original, cumpliendo funciones de mercado diferentes y no un mero sustituto para el material protegido. Por el contrario, de constatarse por la parte actora una finalidad sustancialmente similar al original en conjunción a la existencia de un uso comercial, resultará complicado encajar el mismo en el concepto de «actividad transformativa», en línea con el reciente asunto Andy Warhol Foundation for the Visual Arts, Inc. versus Goldsmith, 598 US (Lin, 2023).

A nivel comunitario, existen diferentes asuntos planteados ante el Tribunal de Justicia de la Unión Europea en los que se han abordado aspectos de los límites al derecho de reproducción, como el asunto Infopaq, por el que se estableció que una actividad realizada en el contexto de un procedimiento de recopilación de datos, por la que se almacena en memoria e imprime un extracto de una obra protegida por el derecho de propiedad intelectual, constituía una reproducción parcial a los efectos del artículo 2 de la Directiva 2001/29/CE si el producto de dicho procedimiento expresaba la creación intelectual del autor (Tribunal de Justicia de la Unión Europea, 2009). No obstante, se requería que el acto fuera provisional y transitorio, que formara parte integrante y esencial de un proceso tecnológico, cuya única finalidad consistiera en facilitar una transmisión en una red entre terceras partes por un intermediario o una utilización lícita, y que dicho acto no tuviera una significación económica independiente (Hugenholtz y Quintais, 2021).

Finalmente, otra de las cuestiones jurídicas más relevantes respecto al *output* de salida se centra en la atribución de titularidad o autoría del mismo y el grado de intervención del usuario en la obtención de ese resultado, tratándose todavía de una cuestión abierta. En territorio estadounidense, algunos planteamientos sostienen que los resultados ofrecidos por el modelo se tratarían de una obra realizada por contrato sobre la base del título 17 del United States Code, capítulo 2, parágrafo 201, perteneciendo al usuario que proporcina el *prompt* de entrada, debiendo rechazar la calificación como autor o coautor a quien simplemente describe a otro de qué modo debería funcionar o verse el trabajo encargado (United States Court Appeals, 1989).

Otros planteamientos, en cambio, consideran que en tanto los modelos pertenecerían a OpenAI, esta sería titular de sus resultados, si bien la misma habría cedido a los usuarios la titularidad sobre los *outputs* conforme a sus condiciones de uso de 31 de enero de 2024. Así, de acuerdo con dichas condiciones y en la medida en que lo permitiera la ley aplicable, el usuario conservaría sus derechos de propiedad sobre el contenido de la entrada y sería propietario de los resultados de salida, asignando OpenAI al usuario todos sus derechos, títulos e intereses, si los hubiere, respecto al *output*.

Otro argumento defiende que los usuarios son los autores de los resultados proporcionados por la inteligencia artificial, dada la intervención activa de los mismos sobre el modelo a través del *prompt*. No obstante, el criterio sostenido por la United States Copyright

Office (USCO) parece ser contrario a este argumento. En su decisión de 14 de febrero de 2022, la USCO rechazó la protección de derechos de autor sobre la obra digital titulada *A Recent Entrance to Paradise*, realizada por un sistema de inteligencia artificial, porque la obra carecía de autoría humana (USCO, 2022). Del mismo modo, en la decisión respecto a la solicitud de Kristina Kashtanova del registro de propiedad intelectual del comic *Zarya of the Dawn*, la USCO consideró que la novela gráfica compuesta por texto escrito por humanos, combinado con imágenes generadas por el servicio de inteligencia artificial Midjourney, constituía una obra protegida por derechos de autor, pero que las imágenes individuales en sí mismas no podían estar protegidas por derechos de autor (USCO, 2023b); concluyendo que la usuaria del sistema no era autora de las imágenes individuales generadas, y ello aun cuando hubiera intervenido a través del empleo de diferentes *prompts*, al no existir ninguna garantía de que una indicación concreta o *prompt* generara un resultado determinado, de forma que las indicaciones o *prompts* funcionarían más como sugerencias que como órdenes, similares a las de un cliente que contrata a un artista para crear una imagen a partir de ciertas instrucciones sobre su contenido.

La Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence de la USCO (2023a) establece un listado de pautas para poder registrar obras que contengan material generado por inteligencia artificial, haciendo hincapié en el requisito de la autoría humana y su contribución –no registrándose obras producidas exclusivamente por máquinas u otros procesos que operen de forma aleatoria o automática, sin ningún aporte o intervención creativa–, pues lo importante será determinar hasta qué punto el ser humano tuvo control creativo sobre la expresión de la obra y hasta qué punto realmente formó los elementos tradicionales de la autoría, siendo tal aspecto lo fundamentalmente protegible. A este respecto, deberá identificarse el grado de aportación humana y la parte generada por inteligencia artificial, no siendo necesario incluir en la solicitud aquella tecnología de inteligencia artificial o empresa que la proporcionó, como tampoco aquel contenido *de minimis*, el cual también quedará excluido de la solicitud (USCO, 2023a).

En la mayoría de los sistemas jurídicos actuales los derechos de autor solo se aplicarán a obras originales –teniendo en cuenta que la originalidad ha de reflejar una creación intelectual propia de un autor humano– y que, por ende, la actividad de creación es intrínsecamente humana, no pudiendo reconocerse a la inteligencia artificial como autora ni como titular de derechos de autor.

Aunque en sistemas jurídicos como el de Reino Unido, Irlanda, Nueva Zelanda, India o Hong Kong se reconoce la autoría de la obra a quien haya tomado las disposiciones necesarias para su creación –de forma que el grado de intervención humana en el proceso constituye un factor determinante–, encontramos algunas distinciones entre obras generadas por una inteligencia artificial sin asistencia humana; obras generadas en colaboración con una inteligencia artificial, en función del grado de contribución humana en la formulación o expresión de ideas; y obras asistidas por una inteligencia artificial en las que la contribución humana es sustancial y la intervención de la inteligencia artificial es mínima.

A nivel comunitario, existe una tendencia favorable a la distinción entre las creaciones humanas asistidas por la inteligencia artificial, esto es, aquellas con intervención y/o dirección humana material, y las creaciones generadas por la inteligencia artificial sin ninguna intervención humana (Comisión Europea, 2020; OMPI, 2020, 2021; Parlamento Europeo, 2020). A fin de constatar si los resultados asistidos por inteligencia artificial pueden protegerse bajo la ley de derechos de autor de la Unión Europea, la Comisión Europea propone una prueba de cuatro pasos: que se trate de una producción del ámbito literario, científico o artístico; que sea el resultado del esfuerzo intelectual humano, reconociendo que es posible crear obras de autoría humana con ayuda de máquinas o dispositivos (Sentencia de 11 de diciembre de 2011, Painer, C-145/10, 2011); que cumpla con los umbrales de originalidad, la cual habrá de ser evaluada por los tribunales nacionales de la Unión Europea y dependerá de si un autor humano ha tomado decisiones creativas durante el proceso de producción, esto es, durante las fases de concepción, ejecución y redacción, y de que estas se reflejen en el resultado final; y que sea identificable con suficiente precisión y objetividad, debiendo concluirse que aquellos resultados generados por inteligencia artificial que carezcan de originalidad, al crearse sin ninguna intervención humana, no cumplen los requisitos para ser considerados como obras susceptibles de protección por derechos de autor.

5. Perspectivas regulatorias a nivel comunitario: breve aproximación a las directivas por responsabilidad civil extracontractual y al reglamento de inteligencia artificial

Entre las principales propuestas regulatorias a nivel comunitario con especial impacto en la implementación de sistemas de inteligencia artificial en el ámbito empresarial y corporativo, destaca la Propuesta de Directiva del Parlamento Europeo y del Consejo relativa a la adaptación de las normas de responsabilidad civil extracontractual a la inteligencia artificial (Directiva sobre responsabilidad en materia de inteligencia artificial), dirigida al establecimiento de requisitos uniformes y a aligerar la carga de la prueba mediante el uso de la exhibición y las presunciones refutables *iuris tantum*.

A tal efecto, establece que, cuando se presente una demanda de responsabilidad por daños y perjuicios que goce de viabilidad por estar respaldada por hechos y pruebas suficientes, los órganos jurisdiccionales podrán ordenar a proveedores, terceros sujetos a sus obligaciones y usuarios la exhibición de pruebas pertinentes relativas a sistemas de inteligencia artificial de alto riesgo específicos de los que se sospeche que han causado daño, en la medida necesaria para sustentar la demanda y con excepción de secretos comerciales e información confidencial. En caso de incumplimiento de la orden de exhibición de la información anterior, el órgano jurisdiccional nacional presumirá el incumplimiento por parte del demandado de un deber de diligencia pertinente. Asimismo, se establece una presunción refutable de relación o nexo de causalidad en caso de culpa, siempre y cuando se constate el incumplimiento de un deber de diligencia, pueda considerarse razonablemente probable

que la culpa ha influido en los resultados producidos por el sistema de inteligencia artificial o en la falta de ellos y el demandante haya demostrado que la información de salida de la inteligencia artificial o la falta de la misma causó los daños.

Otra iniciativa es la Propuesta de Directiva del Parlamento Europeo y del Consejo sobre responsabilidad como consecuencia de los daños causados por productos defectuosos, dirigida a incluir los sistemas de inteligencia artificial en la nueva definición de «producto», reformulando los conceptos «defectuoso», «daño» o «productor» y facilitando la carga de la prueba del demandante. Así, bajo el concepto de «defectuoso» se incluiría no solo la inadecuación al uso del producto, sino también la falta de garantía de seguridad; mientras que el concepto de «daño» incluiría aquellos a la salud psicológica comprobados médicamente y la pérdida o corrupción de datos que no se utilicen exclusivamente con fines profesionales. Por su parte, el nuevo concepto de «productor» sustituye la definición tradicional por un listado amplio de operadores económicos. Asimismo, el demandante solo deberá acreditar el daño, el carácter defectuoso del producto y el nexo causal, sin que sea necesario acreditar la culpabilidad.

Se presumirá el carácter defectuoso cuando pueda probarse que el producto no cumple los requisitos de seguridad establecidos en el derecho de la Unión Europea o en la legislación nacional cuando se demuestre que el daño fue causado por un mal funcionamiento del producto durante su uso o en circunstancias normales o cuando el demandado no cumpla con una orden de exhibición de pruebas. Asimismo, se presumirá el nexo causal entre el carácter defectuoso del producto y el daño cuando se constate que el producto era defectuoso y el daño causado fuera un daño compatible normalmente con el defecto en cuestión. Ahora bien, cuando el tribunal considere que el demandante se enfrenta a dificultades técnicas o científicas excesivas para acreditar el carácter defectuoso del producto, el nexo causal, o ambas cosas, podrá establecer una presunción *iuris tantum* si el demandante demuestra que el producto contribuyó a daños o sea probable que el producto sea defectuoso o que su carácter defectuoso fue una causa probable de los daños, o ambos (Etreros y Sánchez, 2022).

Sin embargo, la iniciativa más importante a este respecto es la relativa a la Propuesta de Reglamento del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial (Ley de inteligencia artificial) y se modifican determinados actos legislativos de la Unión, dirigida a establecer un marco jurídico uniforme en cuanto al control, monitorización, introducción en el mercado y utilización de sistemas de inteligencia artificial en la Unión Europea a partir de un enfoque basado en los riesgos. Este régimen alcanzará a todos los participantes en la cadena de valor de los sistemas de inteligencia artificial incluidos dentro de su ámbito de cobertura, incluyendo, entre otros, a proveedores –lo que incluirá a quienes desarrollen sistemas de inteligencia artificial, como aquellas entidades para las que se desarrollen tales sistemas con vistas a introducirlos en el mercado o ponerlos en servicio con su propio nombre o marca comercial, ya sea de manera remunerada o gratuita–, usuarios –referidos a toda persona física o jurídica, autoridad

pública, agencia u organismo de otra índole que utilice un sistema de inteligencia artificial bajo su propia autoridad, salvo cuando su uso se enmarque en una actividad personal de carácter no profesional–, así como a implementadores o distribuidores, entre otros. Los importadores que introduzcan en el mercado o pongan en servicio un sistema de inteligencia artificial que lleve el nombre o la marca comercial de una persona física o jurídica establecida fuera de la Unión tendrán que garantizar que el proveedor extranjero ya haya efectuado el procedimiento adecuado de evaluación, lleve un marcado europeo de conformidad (CE) y vaya acompañado de la documentación y de las instrucciones de uso necesarias.

Del mismo modo, estarán previstas determinadas obligaciones para los proveedores de modelos de inteligencia artificial de uso general, incluidos los grandes modelos generativos de inteligencia artificial, lo que afectará a la implementación de modelos de OpenAI. La propuesta propone un planteamiento basado en el riesgo que consta de cuatro niveles. Así, diferencia entre «sistemas de inteligencia artificial de riesgo bajo o mínimo», que podrán desarrollarse y utilizarse con arreglo a la legislación vigente, sin obligaciones jurídicas adicionales; «sistemas de inteligencia artificial de alto riesgo», que tienen un impacto potencial negativo en la seguridad de las personas o en sus derechos fundamentales y que para ser permitidos deberán cumplir con un extenso listado de obligaciones, incluida una evaluación *ex ante*; «sistemas de inteligencia artificial de riesgo inaceptable», que suponen la vulneración de derechos fundamentales y, en consecuencia, se considerarán prohibidos; y «sistemas de inteligencia artificial de riesgo específico para la transparencia», respecto de los cuales se imponen obligaciones específicas. Asimismo, aborda los denominados «sistemas de inteligencia artificial de propósito general», así como los «modelos fundacionales», los cuales deberán cumplir obligaciones específicas de transparencia y cumplimiento en materia de propiedad intelectual durante su entrenamiento, entre otros, antes de ser introducidos en el mercado.

La propuesta establece un régimen específico para aquellos supuestos de riesgo sistémico, el cual podría surgir de los modelos de inteligencia artificial de propósito general, incluidos los grandes modelos generativos de inteligencia artificial, respecto de los cuales se prevé la imposición de obligaciones más estrictas. De acuerdo con la Comisión, por ahora, se considera que los modelos de inteligencia artificial de propósito general que se entrenaron utilizando una potencia de cálculo total de más de 10^{25} FLOP (*floating point operations per second*) conllevan riesgos sistémicos, lo que incluiría sistemas como GPT-4 de OpenAI.

6. Conclusiones

El presente estudio proporciona una aproximación a la arquitectura, al funcionamiento, a las aplicaciones, a las limitaciones y a los riesgos asociados a la implementación en el sector jurídico de ChatGPT, así como de aquellas soluciones API basadas en la familia de modelos GPT, entre los que se incluye GPT-4. La comprensión de sus características, del corpus de entrenamiento y de las diferentes fases de desarrollo, preentrenamiento y ajuste por refuerzo

con supervisión humana expuestas en el presente estudio permiten comprender algunas de las principales funcionalidades, limitaciones y riesgos. El creciente interés por este tipo de sistemas advierte de un potencial cambio de paradigma en la concepción de la profesión jurídica, dadas las amplias aplicaciones de estos sistemas. Sin embargo, a pesar de sus utilidades, los modelos continúan presentando importantes limitaciones y riesgos, no solo técnicos, sino también jurídicos y éticos. Frente a la existencia de alucinaciones y sesgos, resultará imprescindible adaptar no solo una batería de medidas técnicas apropiadas, sino también realizar una aproximación adecuada y responsable desde el punto de vista de los usuarios. A nivel jurídico, algunos de los principales riesgos inherentes a los sistemas como ChatGPT y a las soluciones basadas en GPT-4 se circunscriben a aspectos relacionados con la protección de datos y la confidencialidad, los derechos de propiedad intelectual e industrial y los derechos de autor y bases de datos, resultando imperativo analizar cada una de tales cuestiones por separado. Finalmente, entre las propuestas a nivel comunitario con impacto en la implementación de sistemas de inteligencia artificial destacan aquellas dirigidas a armonizar la regulación en la Unión Europea, con exponentes en materia de responsabilidad extracontractual, productos defectuosos y, especialmente, a través de la propuesta de Ley de inteligencia artificial.

Referencias bibliográficas

Accenture. (2021). *Research Based on Analysis of Occupational Information Network*.

Adams, K. (2022). *ChatGPT Won't Fix Contracts.* Adam on Contract Drafting. https://www.adamsdrafting.com/chatgpt-wont-fix-contracts/

Addams, G., Fabbri, A., Ladhak, F., Lehman, E. y Elhadad, N. (2023). *From Sparse to Dense: GPT-4 Summarization with Chain of Density Prompting.* arXiv. https://arxiv.org/abs/2309.04269

Adlakha, V., BehnamGhader, P., Han Lu, X., Meade, N. y Reddy, S. (2023). *Evaluating Correctness and Faithfulness of Instruction-Following Models for Question Answering*. arXiv. https://arxiv.org/pdf/2307.16877.pdf

AEPD. (2018). *Informe del Gabinete Jurídico AEPD 181/2018 (N/REF: 210070/2018)*. https://www.aepd.es/documento/2018-0181.pdf

AEPD. (2020). *Adecuación al RGPD de tratamientos que incorporan inteligencia artificial. Una introducción*. https://www.aepd.es/documento/adecuacion-rgpd-ia.pdf

AEPD. (2021a). *Informe del Gabinete Jurídico AEPD 81/2019 (N/REF: 028891/2019).* https://www.aepd.es/documento/2019-0081.pdf

AEPD. (2021b). *Informe del Gabinete Jurídico AEPD 89/2020 (N/REF: 0089/2020).* https://www.aepd.es/documento/2020-0089.pdf

AEPD. (2023a). *Informe del Gabinete Jurídico AEPD 52/2023 (N/REF: 0052/2023)*. https://www.aepd.es/documento/2023-0052.pdf

AEPD. (2023b). *Inteligencia artificial: sistema vs. tratamiento, medios vs. finalidad*. https://www.aepd.es/prensa-y-comunicacion/blog/inteligencia-artificial-sistema-vs-tratamiento-medio-vs-finalidad

Agencia Tributaria. (2020). *Plan estratégico de la Agencia Tributaria 2020-2023.*

Aletras, N., Androutsopoulos, I., Barrett, L. y Preoţiuc-Pietro, D. (Eds.). (2020). Natural legal language processing workshop 2020. *CEUR Workshop Proceedings*, *2.645*.

Aletras, N., Ash, E., Barrett, L., Chen, D., Meyers, A., Preoţiuc-Pietro, D., Rosenberg, D. y Stent, A. (Eds.). (2019). *Natural Legal Language Processing (NLLP). Proceedings of the 2019 Workshop*. Association for Computational Linguistics. https://aclanthology.org/W19-22.pdf

Aletras, N., Tsarapatsanis, D., Preoţiuc-Pietro, D. y Lampos, V. (2016). Predicting judicial decisions of the European Court of Human Rights: a natural language processing perspective. *PeerJ Computer Science*, *2*(2), 1-19.

Allen & Overy. (2023). *A&O Announces Exclusive Launch Partnership with Harvey.* https://www.allenovery.com/en-gb/global/news-and-insights/news/ao-announces-exclusive-launch-partnership-with-harvey

Ambrogi, B. (2023). *New GPT-Based Chat App from LawDroid is a Lawyer's «Copilot» for Research, Drafting, Brainstorming and More.*

Arts, S., Hou, J. y Gomez, J. C. (2021). Natural language processing to identify the creation and impact of new technologies in patent text: code, data, and new measures. *Research Policy*, *50*(2), 1-13. https://doi.org/10.1016/j.respol.2020.104144

Bacas, T. (2022). *ANALYSIS: Will ChatGPT Bring AI to Law Firms? Not Anytime Soon*. Bloomberg Law. https://news.bloomberglaw.com/bloomberg-law-analysis/analysis-will-chatgpt-bring-ai-to-law-firms-not-anytime-soon

Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., Joseph, N., Kadavath, S., Kernion, J., Conerly, T., El-Showk, S., Elhage, N., Hatfield-Dodds, Z., Hernandez, D., Hume, T., … y Kaplan, J. (2022). *Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback*. arXiv. https://arxiv.org/pdf/2204.05862.pdf

Beltagy, I., Peters, M. E. y Cohan, A. (2020). *Longformer: The Long-Document Transformer.* arXiv. https://arxiv.org/pdf/2004.05150.pdf

Bender, E. M. y Friedman, B. (2018). Data statements for natural language processing: toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, *6*, 587-604.

Bhaskar, A., Fabbri, A. y Durrett, G. (2023). *Prompted Opinion Summarization with GPT-3.5.* https://aclanthology.org/2023.findings-acl.591

Bhattacharya, P., Hiware, K., Rajgaria, S., Pochhi, N., Ghosh, K. y Ghosh, S. (2019). A comparative study of summarization algorithms applied to legal case judgments. En L. S. Azzopardi, B. Stein, N. Fuhr, P. Mayr, C. Hauff y D. Hiemstra (Eds.), *Advances in Information Retrieval (ECIR)*, *11.437*, 413-428.

Bhattacharya, P., Poddar, S., Rudra, K. y Ghosh, K. (2021). Incorporating domain knowledge for extractive summarization of legal case documents. *ICAIL '21. Proceedings of the 18th International Conference on Artificial Intelligence and Law*. arXiv. https://arxiv.org/pdf/2106.15876.pdf

Bommarito, M. J., Martin Katz, D. y Detterman, E. M. (2018). *Lexnlp: Natural Language Processing and Information Extraction for Legal and Regulatory Texts*. arXiv. https://arxiv.org/pdf/1806.03688.pdf

Bommasani, R., Hudson, D., Adeli, E., Altman, R., Arora, S., Arx, S. von, Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Quincy Davis, J., Demszky, D., … y Liang, P. (2022). *On the Opportunities and Risks of Foundation Models*. arXiv. https://arxiv.org/pdf/2108.07258.pdf

Bowman, S. R. (2023). *Eight Things to Know about Large Language Models.* arXiv. https://arxiv.org/abs/2304.00612

Branting, L. K., Pfeifer, C., Brown, B., Ferro, L., Aberdeen, J., Weiss, B., Pfaff, M. y Liao, B. (2021). Scalable and explainable legal prediction. *Artificial Intelligence Law*, *29*, 213-238.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., … y Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems 33 (NeurIPS 2020).* Vancouver, Canadá.

Bruno, A., Mazzeo, P. L., Chetouani, A., Tliba, M. y Kerkouri, M. A. (2023). *Insights into Classifying and Mitigating LLMs' Hallucinations*. arXiv. https://arxiv.org/pdf/2311.08117.pdf

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M. T. y Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early Experiments with GPT-4*. arXiv. https://arxiv.org/pdf/2303.12712.pdf

Burns, C., Ye, H., Klein, D. y Steinhardt, J. (2022). *Discovering Latent Knowledge in Language Models without Supervision.* arXiv. https://arxiv.org/pdf/2212.03827.pdf

Cao, Z., Wei, F., Li, W. y Li, S. (2017). *Faithful to the Original: Fact Aware Neural Abstractive Summarization*. arXiv. https://arxiv.org/pdf/1711.04434.pdf

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M. y Floridi, L. (2018). Artificial intelligence and the «good society»: the US, EU, and UK approach. *Science and Engineering Ethics*, *24*, 505-528. https://doi.org/10.1007/s11948-017-9901-7

CBS News. (2023). *Lawyers Fined for Filing Bogus Case Law Created by ChatGPT*. https://www.cbsnews.com/news/chatgpt-judge-fines-lawyers-who-used-ai/

Cerullo, M. (2023a). *A Lawyer Used ChatGPT to Prepare a Court Filing. It Went Horribly Awry*. CBS News. https://www.cbsnews.com/news/lawyer-chatgpt-court-filing-avianca/

Cerullo, M. (2023b). *Texas Judge Bans Filings Solely Created by AI after ChatGPT Made Up Cases*. CBS News. https://www.cbsnews.com/news/texas-judge-bans-chatgpt-court-filing/

Chalkidis, I., Androutsopoulos, I. y Aletras, N. (2019). Neural legal judgment prediction in English. En A. Korhonen, D. Traum y L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4.317-4.323). Association for Computational Linguistics.

Chalkidis, I., Androutsopoulos, I. y Michos, A. (2017). Extracting contract elements. *Proceedings of the 16th Edition of the International Conference on Articial Intelligence and Law* (pp. 19-28).

Chalkidis, I., Androutsopoulos, I. y Michos, A. (2018). Obligation and prohibition extraction using hierarchical RNNs. En I. Gurevych y Y. Miyao (Eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Vol. 2, Short Papers, pp. 254-259). Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, M., Kotitsas, S., Malakasiotis, P., Aletras, N. y Androutsopoulos, I. (2020). An empirical study on large-scale multi-label text classification including few and zero-shot labels. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7.503-7.515). Association for Computational Linguistics.

Chalkidis, I., Fergadiotis, M., Malakasiotis, P., Aletras, N. y Androutsopoulos, I. (2020). *LEGAL-BERT: The Muppets Straight out of Law School*. arXiv. https://arxiv.org/pdf/2010.02559.pdf

Chalkidis, I., Fergadiotis, M., Tsarapatsanis, D., Aletras, N., Androutsopoulos, I. y Malakasiotis, P. (2021). Paragraph-level rationale extraction through regularization: a case study on European Court of Human Rights Cases. En K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, R. Cotterell, T. Chakraboty e Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 226-241). Association for Computational Linguistics.

Chalkidis, I., Jana, A., Hartung, D., Bommaritto, M., Androutsopoulos, I., Martin Katz, D. y Aletras, N. (2022). *LexGLUE: A Benchmark Dataset for Legal Language Understanding in English.* arXiv. https://arxiv.org/pdf/2110.00976v4.pdf

Chalkidis, I. y Kampas, D. (2019). Deep learning in law: early adaptation and legal word embeddings trained on large corpora. *Artificial Intelligence and Law*, *27*, 171-198.

Chan, l., Garriga-Alonso, A., Goldowsky-Dill, N., Greenblatt, R., Nitishinskaya, J., Radhakrishnan, A. y Shlegeris, B. (2022). *Causal Scrubbing: A Method for Rigorously Testing Interpretability Hypotheses [Redwood Research].*

Chen, X., Li, M., Gao, X. y Zhang, X. (2022). Towards improving faithfulness in abstractive summarization. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)* (pp. 1-13).

Chen, Y., Sun, Y., Yang, Z. y Lin, H. (2020). Joint entity and relation extraction for legal documents with legal feature enhancement. En D. Scott, N. Bel y C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 1.561-1.571). Association for Computing Machinery.

Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Won Chung, H., Sutton, C., Gehrmann, S., Schuh, P., Shi, K., Tsvyashchenko, S., Maynez, J., Rao, A., Barnes, P., Tay, Y., Shazeer, N., Prabhakaran, V., … y Fiedel, N. (2022). *PaLM: Scaling Language Modeling with Pathways, Google Research.* arXiv. https://arxiv.org/pdf/2204.02311.pdf

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kalser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C. y Schulman, J. (2021). *Training Verifiers to Solve Math Word Problems*. arXiv. https://arxiv.org/pdf/2110.14168.pdf

Comisión Europea. (2020). *Trends and Developments in Artificial Intelligence*. https://ec.europa.eu/newsroom/dae/redirection/document/71193

Cuatrecasas. (2023). *Cuatrecasas sella una alianza estratégica con Harvey para implantar la IA generativa*. https://www.cuatrecasas.com/es/spain/art/cuatrecasas-sella-una-alianza-estrategica-con-harvey-para-implantar-la-ia-generativa

Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V. y Salakhutdinov, R. (2019). *Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context*. arXiv. https://arxiv.org/pdf/1901.02860.pdf

Dev, S. y Phillips, J. (2019). *Attenuating Bias in Word Vectors*. arXiv. https://arxiv.org/pdf/1901.07656.pdf

Devlin, J., Chang, M.-W., Lee, K. y Toutanova, K. (2019). *BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://arxiv.org/pdf/1810.04805.pdf

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., DasSarma, N., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Jones, A., Kernion, J., Lovitt, L., Ndousse, K., … y Olah, C. (2021). A mathematical framework for transformer circuits.

Anthropic. https://transformer-circuits.pub/2021/framework/index.html

Etreros, J. y Sánchez, R. (2022). *Responsabilidad civil e inteligencia artificial*. Economic & Jurist. https://www.economistjurist.es/articulos-juridicos-destacados/responsabilidad-civil-e-inteligencia-artificial/

Expert.AI. (2023). *Cuatrecasas incorpora la inteligencia artificial a sus procesos de trabajo*. https://www.expert.ai/es/cuatrecasas-incorpora-la-inteligencia-artificial-a-sus-procesos-de-trabajo/

Fernandes, P., Madaan, A., Lin, E., Farinhas, A., Martins, P. H., Bertsch, A., Souza, J. G. C. de, Zhou, S., Wu, T., Neubig, G. y Martins, A. F. T. (2023). *Bridging the Gap: A Survey on Integrating (Human) Feedback for Natural Language Generation*. arXiv. https://arxiv.org/pdf/2305.00955.pdf

Ferrara, E. (2023). *Should Chatgpt Be Biased? Challenges and Risks of Bias in Large Language Models*. arXiv. https://arxiv.org/pdf/2304.03738.pdf

Ferro, L., Aberdeen, J., Branting, K., Pfeifer, C., Yeh, A. y Chakraborty, A. (2019). Scalable methods for annotating legal-decision corpora. En N. Aletras, E. Ash, L. Barrett, D. Chen, A. Meyers, D. Preotiuc-Prieto, D. Rosenber y A. Stent (Eds.), *Proceedings of the Natural Legal Language Processing Workshop* (pp. 12-20). Association for Computational Linguistics.

Fortune Business Insights. (2023). *AI Market Size Report.* https://www.fortunebusinessinsights.com/industry-reports/artificial-intelligence-market-100114

Galgani, F., Compton, P. y Hoffmann, A. (2012). Towards automatic generation of catch-phrases for legal case reports. *International Conference on Intelligent Text Processing and Computational Linguistics* (pp. 414-425).

Gao, X., Singh, M. P. y Mehra, P. (2012). Mining business contracts for service exceptions. *IEEE Transactions on Services Computing*, *5*(3), 333-344. IEEE.

García Vidal, Á. (2020). Propiedad intelectual y minería de textos y datos: estudio de los artículos 3 y 4 de la Directiva (UE) 2019/790. *Actas de Derecho Industrial y Derecho de Autor*, *40 (2019-2020)* (pp. 99-124). Universidad de Santiago de Compostela.

George, C. y Stuhlmüller, A. (2023). *Factored Verification: Detecting and Reducing Hallucination in Summaries of Academic Papers*. arXiv. https://arxiv.org/pdf/2310.10627.pdf

Goyal, T., Li, J. J. y Durrett, G. (2023). *News Summarization and Evaluation in the Era of GPT-3.* arXiv. https://arxiv.org/abs/2209.12356

Grand View Research. (2023). *Artificial Intelligence Market Size*. https://www.grandviewresearch.com/industry-analysis/artificial-intelligence-ai-market

Guan, J., Dodge, J., Wadden, D., Huang, M. y Peng, H. (2023). *Language Models Hallucinate, but May Excel at Fact Verification*. arXiv. https://arxiv.org/pdf/2310.14564.pdf

Guo, Z., Schlichtkrull, M. y Vlachos, A. (2022). A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, *10*, 178-206.

Han, X., Zhang, Z., Ding, N., Gu, Y., Liu, X., Huo, Y., Qiu, J., Zhang, A., Zhang, L., Han, W., Huang, M., Jin, Q., Lan, Y., Liu, Y., Liu, Z., Lu, Z., Qiu, X., Song, R., Tang, J., … y Zhu, J. (2021). Pre-trained models: past, present and future. *AI Open*, *2*, 225-250.

Hatzius, J., Briggs, J., Kodnani, D. y Pierdomenico, G. (2023). *The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani).* Goldman Sachs. https://www.ansa.it/documents/1680080409454_ert.pdf

Hegel, A., Shah, M., Peaslee, G., Roof, B. y Elwany, E. (2021). *The Law of Large Docu-*

ments: Understanding the Structure of Legal Contracts Using Visual Cues. arXiv. https://arxiv.org/pdf/2107.08128.pdf

Henderson, P., Li, X., Jurafsky, D., Hashimoto, T., Lemley, M. A. y Liang, P. (2023). *Foundation Models and Fair Use*. arXiv. https://arxiv.org/pdf/2303.15715.pdf

Hendrycks, D., Burns, C., Chen, A. y Ball, S. (2021). CUAD: an expert-annotated NLP dataset for legal contract review. *35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1).* arXiv. https://arxiv.org/pdf/2103.06268.pdf

Hu, Z., Li, X., Liu, Z. y Sun, M. (2017). Few-shot charge prediction with discriminative legal attributes. En E. M. Bender, L. Derczynski y P. Isabelle (Eds.), *Proceedings of the 27th International Conference on Computational Linguistics* (pp. 487-498). Association for Computational Linguistics.

Huang, J. y Chang, K. C.-C. (2023). Towards reasoning in large language models: a survey. *Findings of the Association for Computational Linguistics: ACL 2023* (pp. 1.049-1.065). https://aclanthology.org/2023.findings-acl.67.pdf

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B. y Liu, T. (2023). *A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions*. arXiv. https://arxiv.org/pdf/2311.05232.pdf

Hugenholtz, P. B. y Quintais, J. P. (2021). Copyright and artificial creation: does EU copyright law protect AI-assisted output? *IIC. International Review of Intellectual Property and Competition Law*, *52*, 1.190-1.216.

IEEE. (2017). *The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. https://standards.ieee.org/wp-content/uploads/import/documents/other/ead1e.pdf

Jackson, P., Al-Kofahi, K., Tyrrell, A. y Vachher, A. (2003). Information extraction from case law and retrieval of prior cases. *Artificial Intelligence*, *150*, 239-290.

Janiesch, C., Zschech, P. y Heinrich, K. (2021). *Machine Learning and Deep Learning*. arXiv. https://arxiv.org/pdf/2104.05314.pdf

Jelinek, A. (2020). *Preguntas frecuentes sobre la sentencia del Tribunal de Justicia de la Unión Europea en el asunto C-311/18-Comisaria de Protección de Datos vs. Facebook Irlanda y Maximillian Schrems*. European Data Protection Board. https://www.aepd.es/documento/faqs-sentencia-schrems-ii-es.pdf

Ji, Z., Lee, N., Frieske,R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Madotto, A. y Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, *55*(12), 1-38. Association for Computing Machinery.

Kalson, Z. (2022). The implications of ChatGPT and artificial intelligence in family law. *Family Lawyer Magazine*. https://familylawyermagazine.com/chatgpt-and-artificial-intelligence-in-family-law/

Kandpal, N., Deng, H., Roberts, A., Wallace, E. y Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. *Proceedings of the 40th International Conference on Machine Learning* (pp. 15.696-15.707). Association for Computing Machinery.

Kang, C. y Choi, J. (2023). *Impact of Co-occurrence on Factual Knowledge of Large Language Models*. arXiv. https://arxiv.org/pdf/2310.08256.pdf

Katz, D. M., Bommarito, M. J. y Blackman, J. (2017). A general approach for predicting the behavior of the Supreme Court of the United States. *PLoS ONE*, *12*(4). https://doi.org/10.1371/journal.pone.0174698

Katz, D. M., Bommarito, M. J., Gao, S. y Arredondo, P. D. (2023). *GPT-4 Passes the Bar*

Exam. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4389233

Kaufman, A. R., Kraft, P. y Sen, M. (2019). Improving supreme court forecasting using boosted decision trees. *Political Analysis*, *27*, 381-387.

Kien, P. M., Nguyen, H. T., Bach, N. X., Tran, V., Nguyen, M. L. y Phuong, T. M. (2020). Answering legal questions by learning neural attentive text representation. En D. Scott, N. Bel y C. Zong (Eds.), *Proceedings of the 28th International Conference on Computational Linguistics* (pp. 988-998). International Committee on Computational Linguistics.

Kojima, T., Gu, S. S., Reid, M., Matsuo, Y. e Iwasawa, Y. (2022). Large language models are zero-shot reasoners. *36th Conference on Neural Information Processing Systems (NeurIPS 2022)*.

Kowsrihawat, K., Vateekul, P. y Boonkwan, P. (2018). Predicting Judicial decisions of criminal cases from Thai Supreme Court using bi-directional GRU with attention mechanism. *5th Asian Conference on Defense Technology (ACDT)* (pp. 50-55). IEEE.

Lee, K., Ippolito, D., Nystrom, A., Zhang, C., Eck, D., Callison-Burch, C. y Carlini, N. (2022). Deduplicating training data makes language models better. En S. Muresan, P. Nakov y A. Villavicencio (Eds.), *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers, pp. 8.424-8.445). Association for Computational Linguistics.

Leivaditi, S., Rossi, J. y Kanoulas, E. (2020). *A Benchmark for Lease Contract Review.* arXiv. https://arxiv.org/pdf/2010.10386.pdf

Li, S., Li, X., Shang, L., Dong, Z., Sun, C., Liu, B., Ji, Z., Jiang, X. y Liu, Q. (2022). How pre-trained language models capture factual knowledge? A causal-inspired analysis. *Findings of the Association for Computational Linguistics: ACL 2022* (pp. 720-1.732). Association for Computational Linguistics.

Li, Y., Li, Z., Zhang, K., Dan, R., Jiang, S. y Zhang, Y. (2023). ChatDoctor: a medical chat model fine-tuned on a Large Language Model Meta-AI (LLaMA) using medical domain knowledge. *Cureus*, *15*(6). https://arxiv.org/ftp/arxiv/papers/2303/2303.14070.pdf

Lin, P. K. (2023). Retrofitting fair use: art & generative AI after Warhol. *Santa Clara Law Review*, *66*, 1-31.

Lin, S., Hilton, J. y Evans, O. (2022). TruthfulQA: measuring how models mimic human falsehoods. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers, pp. 3.214-3.252). Association for Computational Linguistics.

Lippi, M., Pałka, P., Contissa, G., Lagioia, F., Micklitz, H.-W. Sartor, G. y Torroni, P. (2019). CLAUDETTE: an automated detector of potentially unfair clauses in online terms of service. *Artificial Intelligence and Law*, *27*(2), 117-139. https://link.springer.com/article/10.1007/s10506-019-09243-2

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. y Zhu, C. (2023). *G-EVAL: NLG Evaluation Using GPT-4 with Better Human Alignment*. arXiv. https://arxiv.org/pdf/2303.16634.pdf

Liu, N. F., Lin, K., Hewitt, J., Paranjape, A., Bevilcqua, M., Petroni, F. y Liang, P. (2023). *Lost in the Middle: How Language Models Use Long Contexts*. arXiv. https://arxiv.org/abs/2307.03172

Locke, D. y Zuccon, G. (2022). Case law retrieval: accomplishments, problems, methods and evaluations in the past 30 years. *ACM Computing Surveys*, *1*(1), 1-37. https://arxiv.org/pdf/2202.07209.pdf

Lomas, N. (2019). *Researchers Spotlight the Lie of «Anonymous» Data*. TechCrunch. https://techcrunch.com/2019/07/24/researchers-spotlight-the-lie-of-anonymous-data/

Long, S., Tu, C., Liu, Z. y Sun. M. (2019). Automatic judgment prediction via legal reading comprehension. En M. Sun, X. Huang, H. Ji, Z. Liu y Y. Liu (Eds.), *Chinese Computational Linguistics* (Vol. 11.856, pp. 558-572).

Lovering, C. y Pavlick, E. (2022). Unit testing for concepts in neural networks. En B. Roark y A. Nenkova (Eds.), *Transactions of the Association for Computational Linguistics*, *10*, 1.193-1.208.

Luo, B., Feng, Y., Xu, J., Zhang, X. y Zhao, D. (2017). Learning to predict charges for criminal cases with legal basis. En M. Palmer, R. Hwa y S. Riedel (Eds.), *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2.727-2.736). Association for Computational Linguistics.

Mallen, A., Asai, A., Zhong, V., Das, R., Khashabi, D. y Hajishirzi, H. (2023). When not to trust language models: investigating effectiveness of parametric and non-parametric memories. En A. Rogers, J. Boyd-Graber y N. Okazaki (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (Vol. 1, Long Papers, pp. 9.802-9.822). Association for Computational Linguistics.

Markovski, Y. (2023). *How Your Data is Used to Improve Model Performance*. OpenAI. https://help.openai.com/en/articles/5722486-how-your-data-is-used-to-improve-model-performance

Maynez, J., Narayan, S., Bohnet, B. y McDonald, R. (2020). *On Faithfulness and Factuality in Abstractive Summarization.* arXiv. https://arxiv.org/pdf/2005.00661.pdf

McCarthy, J., Minsky, M. L., Rochester, N. y Shannon, C. E. (1955). *A proposal for the Dartmouth Summer Research Project on Artificial Intelligence.* Stanford University. http://jmc.stanford.edu/articles/dartmouth/dartmouth.pdf

McKenna, N., Li, T., Cheng, L., Hosseini, M. J., Johnson, M. y Steedman, M. (2023). *Sources of Hallucination by Large Language Models on Inference Tasks*. arXiv. https://arxiv.org/pdf/2305.14552.pdf

Medvedeva, M., Vols, M. y Wieling, M. (2018). Judicial decisions of the European Court of Human Rights: looking into the crystal ball. *Proceedings of the Conference on Empirical Legal Studies in Europe 2018* (pp. 1-24). https://martijnwieling.nl/files/Medvedeva-submitted.pdf

Medvedeva, M., Vols, M. y Wieling, M. (2020). Using machine learning to predict decisions of the European Court of Human Rights. *Artificial Intelligence and Law*, *28*(2), 237-266.

Mencia, E. L. y Furnkranzand, J. (2010). Efficient multilabel classification algorithms for large-scale problems in the legal domain. En E. Francesconi, S. Montemagni, W. Peters y D. Tiscornia (Eds.), *Semantic Processing of Legal Texts, Lecture Notes in Computer Science* (Vol. 6.036, pp. 192-215). Springer.

Meng, K., Sharma, A., Andonian, A., Beclinkov, Y. y Bau, D. (2023). *Mass-Editing Memory in a Transformer*. arXiv. https://arxiv.org/pdf/2210.07229.pdf

Merken, S. (2023). *New York Lawyers Sanctioned For Using Fake ChatGPT Cases in Legal Brief.* Reuters. https://www.reuters.com/legal/new-york-lawyers-sanctioned-using-fake-chatgpt-cases-legal-brief-2023-06-22/

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.-T., Ko, P. W., Iyyer, M., Zettlemoyer, L. y Hajishirzi, H. (2023). *FACTSCORE: Fine-Grained Atomic Evaluation of Factual Precision in Long Form Text Generation*. https://arxiv.org/pdf/2305.14251.pdf

Moore, P. V. (2023). *Inteligencia artificial en el entorno laboral. Desafíos para los trabaja-*

dores. OpenMind BBVA. https://www.bbvaopenmind.com/articulos/inteligencia-artificial-en-entorno-laboral-desafios-para-trabajadores/

Mumcuoğlu, E., Öztürk, C. E. y Ozaktas, H. M. (2021). Natural language processing in law: prediction of outcomes in the higher courts of Turkey. *Information Processing & Management*, *58*(5). https://doi.org/10.1016/j.ipm.2021.102684

Nallapati, R. y Manning, C. D. (2008). Legal docket-entry classification: where machine learning stumbles. *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing* (pp. 438-446). Association for Computational Linguistics.

Navarro, E. (2023). *How can ChatGPT impact legal services?* Consejo General de la Abogacía Española.

Niklaus, J., Chalkidis, I. y Stürmer, M. (2021). *Swiss-Judgment-Prediction: A Multilingual Legal Judgment Prediction Benchmark*. arXiv. https://arxiv.org/pdf/2110.00806.pdf

Nye, M., Andreassen, A. J., Gur-Ari, G., Michalewski, H., Austin, J., Bieber, D., Dohan, D., Lewkowycz, A., Bosma, M., Luan, D., Sutton, C. y Odena, A. (2022). *Show Your Work: Scratchpads for Intermediate Computation with Language Models*. arXiv. https://arxiv.org/pdf/2112.00114.pdf

OCDE. (2019). *Recommendation of the Council on OECD Legal Instruments Artificial Intelligence*.

OMPI. (2020). *Versión revisada del documento temático sobre las políticas de propiedad intelectual y la inteligencia artificial*.

OMPI. (2021). *WIPO Conversation on Intellectual Property (IP) and Artificial Intelligence (AI): Third Session*.

Onoe, Y., Zhang, M., Choi, E. y Durrett, G. (2022). Entity cloze by date: what LMs know about unseen entities. En M. Carpuat, M.-C. de Marneffe e I. V. Meza Ruiz (Eds.), *Findings of the Association for Computational Linguistics: NAACL 2022* (pp. 693-702).

OpenAI. (s. f.). *OpenAI Personal Data Removal Request*.

OpenAI. (2019). *Request for Comment on Intellectual Property Protection for Artificial Intelligence Innovation, PTO-C-2019-0038.* United States Patent and Trademark Office. Department of Commerce.

OpenAI. (2023a). *Condiciones de uso*. https://openai.com/policies/terms-of-use

OpenAI. (2023b). *Custom Instructions for ChatGPT*. https://openai.com/blog/custom-instructions-for-chatgpt

OpenAI. (2023c). *GPT-4 Technical Report.* arXiv. https://arxiv.org/pdf/2303.08774.pdf

OpenAI. (2023d). *Política de privacidad*. https://openai.com/policies/privacy-policy

OpenAI. (2023e). *Política de privacidad para la UE*. https://openai.com/es/policies/eu-privacy-policy

OpenAI. (2024a). *Enterprise Privacy at OpenAI*. https://openai.com/enterprise-privacy

OpenAI. (2024b). *Usage Policies*. https://openai.com/policies/usage-policies

OpenAI. (2024c). *How ChatGPT and Our Language Models Are Developed*. https://help.openai.com/en/articles/7842364-how-chatgpt-and-our-language-models-are-developed

OpenAI. (2024d). *OpenAI Personal Data Removal Request*. https://share.hsforms.com/1UPy6xqxZSEqTrGDh4ywo_g4sk30

OpenAI. (2024e). *OpenAI Privacy Request Portal.* https://privacy.openai.com/policies?name=open-ai-privacy-request-portal#privacy-practices

OpenAI. (2024f). *Data Processing Addendum*. https://openai.com/policies/data-processing-addendum

Ouyang, L., Wu, J., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J. y Lowe, R. (2022). *Training Language Models to Follow Instructions with Human Feedback.* arXiv. https://arxiv.org/pdf/2203.02155.pdf

Parlamento Europeo. (2020). *Resolución del Parlamento Europeo, de 20 de octubre de 2020, sobre los derechos de propiedad intelectual para el desarrollo de las tecnologías relativas a la inteligencia artificial*. https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_ES.html

Patil, V., Hase, P. y Bansal, M. (2023). *Can Sensitive Information Be Deleted from LLMs? Objectives for Defending Against Extraction Attacks*. arXiv. https://arxiv.org/pdf/2309.17410.pdf

Perlman, A. (2023). *The Implications of ChatGPT for Legal Services and Society.* Center on the Legal Profession. Harvard Law School. https://clp.law.harvard.edu/knowledge-hub/magazine/issues/generative-ai-in-the-legal-profession/the-implications-of-chatgpt-for-legal-services-and-society/

Pu, D. y Demberg, V. (2023). ChatGPT vs. human-authored text: insights into controllable text summarization and sentence style transfer. En V. Padmakumar, G. Vallejo y Y. Fu (Eds.), *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* (pp. 1-18). Association for Computational Linguistic. https://aclanthology.org/2023.acl-srw.1/

PwC. (2023). *PwC Announces Strategic Alliance with Harvey, Positioning PWC's Legal Business Solutions at the Forefront of Legal Generative AI.* https://www.pwc.com/gx/en/news-room/press-releases/2023/pwc-announces-strategic-alliance-with-harvey-positioning-pwcs-legal-business-solutions-at-the-forefront-of-legal-generative-ai.html

Radford, A., Wu, J., Child, R., Amodei, D. y Sutskever, I. (2019). *Language Models are Unsupervised Multitask Learners*. https://insightcivic.s3.us-east-1.amazonaws.com/language-models.pdf

Rajani, N. F., McCann, B., Xiong, C. y Socher, R. (2019). Explain yourself! Leveraging language models for commonsense reasoning. En A. Korhonen, D. Traum y L. Màrquez (Eds.), *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4.932-4.942).

Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., Smith-Loud, J., Theron, D. y Barners, P. (2020). *Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing*. arXiv. https://arxiv.org/pdf/2001.00973.pdf

Ravichander, A., Black, A. W., Wilson, S., Norton, T. y Sadeh, N. (2019). Question answering for privacy policies: combining computational and legal perspectives. En K. Iniu, J. Jiang, V. Ng y X. Wan (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 4.947-4.958). Association for Computational Linguistics.

Rincón, G. (2023). *El uso de la inteligencia artificial por la Administración Tributaria: ¿quién vigila a los vigilantes?* Garrigues. https://www.garrigues.com/es_ES/garrigues-digital/uso-inteligencia-artificial-administracion-tributaria-quien-vigila-vigilantes

Roberts, G. (2022). *AI Training Datasets: The Books1+Books2 that Big AI Eats for Breakfast*. Vision of Freedom. https://gregoreite.com/drilling-down-details-on-the-ai-training-datasets/

Ruger, T. W., Kim, P. T., Martin, A. D. y Quinn, K. M. (2004). The Supreme Court forecasting project: legal and political science approaches to Supreme Court decision-making. *Columbia Law Review*, *104*(4), 1.150-1.210.

Sánchez, L. (2023). *Francesc Muñoz: «Estoy convencido de que la IA Generativa hará a los abogados mejores»*. Economist & Jurist. https://www.economistjurist.es/zbloque-1/francesc-munoz-estoy-convencido-de-que-la-ia-generativa-hara-a-los-abogados-mejores/

Sánchez Aristi, R., Pérez Marcilla, M. y Andoni Eguiluz, J. (2023). *El desarrollo de sistemas de inteligencia artificial y la posible infracción de derechos de autor*. Cuatrecasas. https://www.cuatrecasas.com/es/spain/art/el-desarrollo-de-sistemas-de-inteligencia-artificial-y-la-posible-infraccion-de-derechos-de-autor

Sartor, G. (2020). *The Impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence.* European Parliamentary Research Service.

Savelka, J., Gray, M. A. y Westermann, H. (2023). *Explaining Legal Concepts with Augmented Large Language Models (GPT-4)*. arXiv. https://arxiv.org/pdf/2306.09525.pdf

Schulman, J., Wolski, F., Dhariwal, P., Radford, A. y Klimov, O. (2017). *Proximal Policy Optimization Algorithms.* arXiv. https://arxiv.org/pdf/1707.06347.pdf

Sellick, M. (2022). *Can AI Replace Patent Attorneys?* HGF. https://www.hgf.com/news/can-ai-replace-patent-attorneys/

Silva, D. de y Alahakoon, D. (2021). *An Artificial Intelligence Life Cycle: From Conception to Production*. arXiv. https://arxiv.org/pdf/2108.13861.pdf

Singhal, K., Tu, T., Gottweis, J., Sayres, R., Wulczyn, E., Hou, L., Clark, K., Pfohl, S., Cole-Lewis, H., Neal, D., Schaekermann, M., Wang, A., Amin, M., Lachgar, S., Mansfield, P., Prakash, S., Green, B., Dominowska, E., Aguera y Arcas, B., … y Natarajan, V. (2023). *Towards Expert-Level Medical Question Answering with Large Language Models*. arXiv. https://arxiv.org/pdf/2305.09617.pdf

Strickson, B. e Iglesia, B. de la. (2020). Legal judgement prediction for UK Courts. *ICISS '20: Proceedings of the 3rd International Conference on Information Science and Systems*. Association for Computing Machinery.

Şulea, O.-M.ª, Zampieri, M., Vela, M. y Genabith, J. van. (2017). Predicting the law area and decisions of french supreme court cases. En R. Mitkov y G. Angelova (Eds.), *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)* (pp. 716-722). Incoma.

Thompson, A. (2022). *What's in my AI?* Life Architect. https://lifearchitect.ai/whats-in-my-ai/

Tiersma, P. M. (1999). *Legal Language*. The University of Chicago Press.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., … y Scialom, T. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv. https://arxiv.org/pdf/2307.09288.pdf

Tran, V., Le Nguyen, M. y Satoh, K. (2019). Building legal case retrieval systems with lexical matching and summarization using a pre-trained phrase scoring model. *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Law, ICAIL '19* (pp. 275-282).

Tribunal de Justicia de la Unión Europea. (16 de julio de 2009). *Infopaq International A/S y Danske Dagblades Forening, C5/08*.

Tuggener, D., Däniken, P. von, Peetz, T. y Cieliebak, M. (2020). LEDGAR: a large-scale multi-label corpus for text classification of legal provisions in contracts. En N. Calzoni, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk y S. Piperidis (Eds.), *Proceedings of the Twelfth Language Resources and Evaluation Conference* (pp. 1.235-1.241). European Language Resources Association.

United States Court Appeals. (13 de septiembre de 1989). *SOS, Inc. v. Payday, Inc.* 886 F.2d 1081 (9th Cir. 1989).

United States Court of Appeals. (6 de febrero de 2002). *Kelly v. Arriba Soft Corp.* 280 F.3d 934 (9th Cir. 2002).

United States District Court. (19 de septiembre de 2023). *Author's Guild v. OpenAI Inc. (1:23-cv-08292)*. Southern District of New York.

Urchs, S., Mitrovic, J. y Granitzer, M. (2021). Design and implementation of german legal decision corpora. En A. P. Rocha, L. Steel y J. van den Herik (Eds.), *Proceedings of the 13th International Conference on Agents and Artificial Intelligence, ICAART* (Vol. 2, pp. 515-521).

USCO. (2022). *Second Request for Reconsideration for Refusal to Register A Recent Entrance to Paradise (Correspondence ID 1-3ZPC6C3; SR # 1-7100387071)*. https://www.copyright.gov/rulings-filings/review-board/docs/a-recent-entrance-to-paradise.pdf

USCO. (2023a). *Copyright Registration Guidance: Works Containing Material Generated by Artificial Intelligence.* Library of Congress.

USCO. (2023b). *Zarya of the Dawn (# VAu001 480196)*. https://www.copyright.gov/docs/zarya-of-the-dawn.pdf

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N. y Kaiser, Ł. (2017). Attention is all you need. *31st Conference on Neural Information Processing Systems (NIPS 2017).* arXiv. https://arxiv.org/pdf/1706.03762.pdf

Virtucion, M. B., Aborot, J. A., Abonita, J. K., Aviñate, R., Copino, R. J. B., Neverida, M. P., Osiana, V. O., Peramo, E. C., Syjuco, J. G. y Tan, G. B. A. (2018). Predicting decisions of the philippine supreme court using natural language processing and machine learning. *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (pp. 130-135). IEEE.

Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E. H., Narang, S., Chowdhery, A. y Zhou, D. (2023). *Self-Consistency Improves Chain of Thought Reasoning in Language Models*. arXiv. https://arxiv.org/pdf/2203.11171.pdf

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Wei Yu, A., Lester, B., Du, N., Dai, A. M. y Le, Q. V. (2022). *Finetuned Language Models are Zero-Shot Learners*. https://openreview.net/pdf?id=gEZrGCozdqR

Wei, J., Wang, X., Schuurman, D., Bosma, M., Ichter, B., Xia, F., Chi, E. H., Le, Q. V. y Zhou, D. (2022). *Chain-of-Thought Prompting Elicits Reasoning in Large Language Models.* arXiv. https://arxiv.org/pdf/2201.11903.pdf

White, J., Fu, Q., Hays, S., Sandborn, M., Olea, C., Gilbert, H., Elnashar, A., Spencer-Smith, J. y Schmidt, D. C. (2023). *A Prompt Pattern Catalog to Enhance Prompt Engineering with ChatGPT*. arXiv. https://arxiv.org/pdf/2302.11382.pdf

White, J., Hays, S., Fu, Q., Spencer-Smith, J. y Schmidt, D. C. (2023). *ChatGPT Prompt Patterns for Improving Code Quality, Refac-*

toring, Requirements Elicitation, and Software Design. arXiv. https://arxiv.org/pdf/2303.07839.pdf

Williams, C. (2005). *Tradition and Change in Legal English. Verbal Constructions in Prescriptive Texts.* Peter Lang Publishing.

World Economic Forum. (2023). *Satya Nadella Says AI Golden Age Is Here and «It's Good for Humanity»*. https://www.weforum.org/press/2023/01/satya-nadella-says-ai-golden-age-is-here-and-it-s-good-for-humanity

Xiao, C., Zhong, H., Guo, Z., Tu, C., Liu, Z., Sun, M., Feng, Y., Han, X., Hu, Z., Wang, H. y Xu, J. (2018). *CAIL2018: A Large-Scale Legal Dataset for Judgment Prediction*. arXiv. https://arxiv.org/pdf/1807.02478.pdf

Xu, C., Sun, Q., Zheng, K., Geng, X., Zhao, P., Feng, J., Tao, C., Lin, Q. y Jiang, D. (2023). *WizardLM: Empowering Large Language Models to Follow Complex Instructions*. arXiv. https://arxiv.org/pdf/2304.12244.pdf

Yang, W., Jia, W., Zhou, X. y Luo, Y. (2019). Legal judgment prediction via multi-perspective bi-feedback network. *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 4.085-4.091).

Ye, H., Jiang, X., Luo, Z. y Chao, W. (2018). Interpretable charge predictions for criminal cases: learning to generate court views from fact descriptions. *Proceedings of NAACL-HLT 2018* (pp. 1.854-1.864). https://aclanthology.org/N18-1168.pdf

Ye, H., Liu, T., Zhang, A., Hua, W. y Jia, W. (2023). *Cognitive Mirage: A Review of Hallucinations in Large Language Models*. arXiv. https://arxiv.org/pdf/2309.06794.pdf

Yu, F., Quartey, L. y Schilder, F. (2022). *Legal Prompting: Teaching a Language Model to Think Like a Lawyer*. https://arxiv.org/pdf/2212.01326.pdf

Zaheer, M., Guruganesh, G., Dubey, A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, Li y Ahmed, A. (2021). Big bird: transformers for longer sequences. *34th Conference on Neural Information Processing Systems* (pp. 17.283-17.297). arXiv. https://arxiv.org/pdf/2007.14062.pdf

Zahn, M. (2023). *Authors' lawsuit against OpenAI Could «Fundamentally Reshape» Artificial Intelligence, According to Experts*. ABC News. https://abcnews.go.com/Technology/authors-lawsuit-openai-fundamentally-reshape-artificial-intelligence-experts/story?id=103379209

Zelikman, E., Wu, Y., Mu, J. y Goodman, N. D. (2022). *STaR: Bootstrapping Reasoning with Reasoning*. https://arxiv.org/abs/2203.14465

Zeng, A., Liu, X., Du, Z., Wang, Z., Lai, H., Ding, M., Yang, Z., Zheng, W., Xia, X., Tam, W. L., Ma, Z., Xue, Y., Zhai, J., Chen, W., Liu, Z., Zhang, P., Dong, Y. y Tang, J. (2023). GLM-130B: an open bilingual pre-trained model. *The Eleventh International Conference on Learning Representations, ICLR 2023*. https://openreview.net/pdf?id=-Aw0rrrPUF

Zhang, S., Dong, L., Li, X., Zhang, S., Sun, X., Wang, S., Li, J., Hu, R., Zhang, T., Wu, F. y Wang, G. (2023). *Instruction Tuning for Large Language Models: A Survey*. https://arxiv.org/pdf/2308.10792.pdf

Zhang, B. H., Lemoine, B. y Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning*. https://arxiv.org/pdf/1801.07593.pdf

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., Chen, Y., Wang, L., Luu, A. T., Bi, W., Shi, F. y Shi, S. (2023). *Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models*. arXiv. https://arxiv.org/pdf/2309.01219.pdf

Zhang, M., Press, O., Merrill, W., Liu, A. y Smith, N. A. (2023). *How Language Model Hallucinations Can Snowball*. arXiv. https://arxiv.org/pdf/2305.13534.pdf

Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X. V., Mihaylov, T., Ott, M., Shleifer, S., Shuster, K., Simig, D., Koura, P. S., Sridhar, A., Wang, T. y Zettlemoyer, L. (2022). *OPT: Open Pre-trained Transformer Language Models*. https://arxiv.org/pdf/2205.01068.pdf

Zheng, S., Huang, J. y Chan, K. C.-C. (2023). *Why Does ChatGPT Fall Short in Providing Truthful Answers?* https://arxiv.org/pdf/2304.10513.pdf

Zhong, H., Guo, Z., Tu, C., Xiao, C., Liu, Z. y Sun, M. (2018). Legal judgment prediction via topological learning. En E. Riloff, D. Chiang, J. Hockenmaier y J. Tsujii (Eds.), *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3.540-3.549). Association for Computational Linguistics.

Zhong, H., Wang, Y., Tu, C., Zhang, T., Liu, Z. y Sun, M. (2020). Iteratively questioning and answering for interpretable legal judgment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(01), 1.250-1.257.

Zhong, H., Xiao, C., Tu, C., Zhang, T., Liu, Z. y Sun, M. (2020). How does NPL benefit legal system: a summary of legal artificial intelligence. En D. Jurafsky, J. Chai, N. Schluter y J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5.218-5230). Association for Computational Linguistics.

Francisco Julio Dosal Gómez. Abogado especialista en derecho internacional de los negocios, arbitraje internacional y derecho internacional de la construcción. Graduado en Derecho por la Universidad de Cantabria (España) y LLM en Derecho Internacional de los Negocios por el Centro de Estudios Garrigues (España). Miembro del Club Español e Iberoamericano del Arbitraje (CEIA) y del Young International Council for Commercial Arbitration (ICCA). En 2023 publicó su artículo titulado «El Dispute Avoidance Adjudication Board en la Rainbow Suite FIDIC de 2017: funcionamiento del sistema de asistencia informal y del sistema de resolución de disputas» en la *Newsletter Dispute Boards* del Club Español e Iberoamericano del Arbitraje (núm. 2, pp. 25-42).

Judith Nieto Galende. Abogada especialista en derecho internacional de los negocios y M&A. Doble grado en Derecho y Administración y Dirección de Empresas por la Universidad Autónoma de Madrid (España) y LLM en Derecho Internacional de los Negocios por el Centro de Estudios Garrigues (España). Miembro de la International Bar Association, del Club Español e Iberoamericano del Arbitraje (CEIA) y del Young International Council for Commercial Arbitration (ICCA). Tras su paso por el área legal de M&A, actualmente trabaja en un fondo de inversiones británico especializado en energías renovables denominado WiseEnergy y cuenta con más de un año de experiencia laboral tanto a nivel nacional como internacional asesorando a clientes en el ámbito legal y financiero.

Contribución de autores. F. J. D. G. y J. N. G. han participado a partes iguales en la elaboración de este estudio de investigación.