

Entrenamiento de inteligencia artificial y límites del derecho de los editores de prensa

* * * *

Iván Vargas-Chaves¹

Universidad Militar Nueva Granada (Colombia)

ivargas@outlook.com

<https://orcid.org/0000-0001-6597-2335>

Recibido: 13 de marzo de 2026

Aceptado: 20 de agosto de 2026

<https://doi.org/10.26422/RIPI.2026.2400.var>

Resumen

Este artículo analiza cómo se relacionan la inteligencia artificial generativa y el derecho de autor en la Unión Europea, centrándose en la petición de decisión prejudicial C-250/25 (*Like Company c. Google*) y en las nuevas obligaciones de transparencia de la Ley de Inteligencia Artificial. Usando un enfoque cualitativo de análisis documental, se revisan documentos procesales, opiniones doctrinales y literatura especializada sobre las excepciones para la minería de textos y datos. Los resultados muestran que el marco legal europeo actual tiene ambigüedades importantes. El derecho conexo de los editores de prensa, regulado por el artículo 15 de la Directiva 2019/790/UE sobre los Derechos de Autor en el Mercado Único Digital, no es suficiente para resolver los retos del entrenamiento de modelos fundacionales y las técnicas de Generación Aumentada por Recuperación. Además, los requisitos de transparencia del nuevo marco europeo sobre inteligencia artificial no resuelven los problemas prácticos de la reserva de derechos (*opt-out*) para los creadores individuales, por ello, aplicar categorías jurídicas tradicionales

1 Profesor Asociado de la Facultad de Derecho “General Luis Carlos Camacho Leyva” de la Universidad Militar Nueva Granada (Bogotá, Colombia). Doctor en Derecho Supranacional e Interno por la Università di Palermo, Italia, y Doctor en Derecho, en la especialidad de Derecho Internacional Privado, por la Universidad de Barcelona, España.

a tecnologías probabilísticas puede frenar la innovación tecnológica. El artículo concluye que, por tanto, es necesario avanzar hacia esquemas más flexibles y justos que permitan el desarrollo de la inteligencia artificial y aseguren una remuneración adecuada para los autores.

Palabras clave: inteligencia artificial generativa, derecho de autor, minería de textos y datos, modelos de lenguaje grande, excepciones de copyright.

Training of Artificial Intelligence, and the Limitations of Press Publishers' Rights

Abstract

This article examines how generative artificial intelligence interacts with copyright law in the European Union. It focuses on the preliminary ruling request C-250/25 (Like Company v. Google) and the new transparency rules in the Artificial Intelligence Act. The study uses a qualitative method, reviewing procedural documents, expert opinions, and literature on text and data mining exceptions. The findings reveal that the current European legal framework is unclear in several areas. The rights of newspaper publishers, as set out in Article 15 of Directive 2019/790/EU, do not fully address the challenges of training foundational models and using Augmented Recall Generation techniques. In addition, the transparency rules in the new AI framework do not solve the practical issues around opt-out rights for individual creators. As a result, using traditional legal categories for these new technologies could slow down innovation. The article suggests that more flexible and fair approaches are needed to support AI development while ensuring creators are properly paid.

Key words: generative artificial intelligence, copyright law, text and data mining, large language model, *copyright* exceptions.

Treinamento de Inteligência Artificial e limites do direito dos editores de imprensa

Resumo

Os resultados demonstram que o atual quadro jurídico europeu apresenta ambiguidades significativas. O direito conexo dos editores de imprensa, regulado pelo artigo 15.º da Diretiva (UE) 2019/790 relativa aos Direitos de Autor no Mercado Único Digital, revela-se insuficiente para resolver os desafios do treinamento de modelos fundacionais e das técnicas de Geração Aumentada por Recuperação (RAG). Além disso, os requisitos de transparência do novo quadro europeu sobre inteligência artificial não resolvem os problemas práticos da reserva de direitos

(*opt-out*) para os criadores individuais; por conseguinte, a aplicação de categorias jurídicas tradicionais a tecnologias probabilísticas pode desestimular a inovação tecnológica. O artigo conclui que é necessário avançar para esquemas mais flexíveis e justos que permitam o desenvolvimento da inteligência artificial e assegurem uma remuneração adequada aos autores.

Palavras-chave: inteligência artificial generativa, direito de autor, mineração de textos e dados, grandes modelos de linguagem, exceções de copyright.

1. Introducción

La adopción masiva de la inteligencia artificial generativa representa uno de los principales desafíos para el derecho de propiedad intelectual en la actualidad. Los modelos de lenguaje grande (LLM, por sus siglas en inglés) pueden asimilar, procesar y generar texto, lo que ha difuminado los límites entre el uso legítimo, la innovación tecnológica y la posible infracción de derechos de autor. En la Unión Europea, este conflicto se está debatiendo en los tribunales, como lo ilustra el caso C-250/25 (*Like Company c. Google Ireland Limited*), que será aquí analizado.

Este artículo se organiza en torno a la siguiente pregunta central: ¿en qué medida el marco normativo europeo actual, enfocado en el derecho conexo de los editores de prensa y las excepciones para la minería de textos y datos, puede resolver la tensión entre la protección del derecho de autor y la innovación en inteligencia artificial generativa sin crear asimetrías de mercado?

Para el análisis, se realizó una búsqueda bibliográfica y documental de fuentes jurídicas especializadas, utilizando palabras clave en inglés y español como *Generative AI*, *Copyright Law*, *Text and Data Mining* (TDM), *Press Publishers' Right* y *AI Act Transparency*. El corpus teórico se definió mediante operadores booleanos en ecuaciones de búsqueda como: (*"Generative AI" OR "Large Language Models"*) AND (*"Copyright" OR "TDM"*) AND (*"European Union" OR "CDSMD"*).

El análisis se fundamentó en la triangulación de información, comparando tres tipos de documentos: 1) procesales y jurisprudenciales, como el texto de remisión del caso C-250/25; 2) opiniones de instituciones y asociaciones especializadas, como la crítica de la European Copyright Society (2026); y 3) literatura académica ac-

tualizada sobre derecho de propiedad intelectual. Esta triangulación permitió reducir sesgos y comparar las limitaciones técnicas del proceso judicial con los avances doctrinales y legislativos.

A lo largo del texto se sostiene la hipótesis de que el marco europeo actual es insuficiente y genera desigualdades estructurales. Se concluye que la dependencia del derecho conexo de los editores (artículo 15 de la Directiva 2019/790/UE sobre los Derechos de Autor en el Mercado Único Digital) y la confianza en las obligaciones de transparencia del Reglamento 2024/1689/UE del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (EU AI Act), no resuelven la vulnerabilidad de los autores originales. Además, la falta de claridad técnica en los tribunales puede limitar la innovación tecnológica y no garantiza una remuneración justa para los creadores.

2. Marco teórico y revisión de literatura

2.1 Fundamentos técnicos y jurídicos de la inteligencia artificial generativa y el uso de datos de entrenamiento

Para comprender las tensiones entre el derecho de autor y la inteligencia artificial generativa, como en el caso C-250/25, es necesario establecer un marco teórico que explique la estructura computacional de estos sistemas y la falta de transparencia en sus datos de entrenamiento. En el centro de esta revolución tecnológica se encuentra el aprendizaje profundo, una rama del aprendizaje automático (Boden, 2017).

El aprendizaje profundo utiliza redes neuronales artificiales organizadas en múltiples capas jerárquicas de nodos computacionales (Goodfellow et al., 2016; Ihalainen, 2018). En el procesamiento del lenguaje natural, este enfoque se basa en la arquitectura Transformer y mecanismos de autoatención, que permiten procesar secuencias textuales en paralelo y captar relaciones sintácticas y semánticas a larga distancia. El corpus de entrenamiento se somete primero a una fase de tokenización, fragmentando el texto en unidades discretas (tokens), que luego se representan como vectores densos en un espacio vectorial multidimensional (Yan et al., 2024). Las capas iniciales

identifican correlaciones estadísticas locales, mientras que las capas intermedias y profundas capturan relaciones contextuales complejas, jerarquías gramaticales e inferencias lógicas a partir de grandes volúmenes de datos (Cotino Hueso y Simón Castellano, 2024).

Durante el preentrenamiento, el modelo ajusta iterativamente cientos de miles de millones de parámetros numéricos, llamados “pesos”, mediante algoritmos de optimización matemática como el descenso de gradiente estocástico. Este proceso busca minimizar la función de pérdida al predecir el siguiente token en una secuencia. Así, desde una perspectiva computacional, más allá de almacenar bases de datos relacionales y copias literales de los datos, estos modelos lo que hacen es codificar representaciones estadísticas abstractas y probabilidades condicionales en su red de parámetros (Brown et al., 2020; Nieva Fenoll, 2018).

Sin embargo, esta premisa técnica se enfrenta al fenómeno de la “memorización”, por el cual, bajo ciertas condiciones de sobreajuste, los modelos generativos pueden recrear copias idénticas o casi idénticas de material protegido utilizado en el entrenamiento. Esta ambivalencia dificulta determinar una posible infracción del derecho de reproducción, ya que sugiere que el modelo retiene la expresión original de forma latente. Los LLM requieren grandes volúmenes de datos; por ejemplo, las primeras versiones de modelos como GPT-3 se entrenaron con bases públicas como Common Crawl, WebText2, bibliotecas completas de libros digitalizados y la base de Wikipedia en inglés (Santarcangelo et al., 2023).

En este contexto, Buick (2025) señala que la mayoría de este material está protegido por derechos de autor, ya que la protección se aplica automáticamente a cualquier obra original y se extiende por muchos años. Aunque existen obras de dominio público o bajo licencias como Creative Commons, estas fuentes no bastan para entrenar modelos avanzados. Además, usarlas exclusivamente podría obligar a los desarrolladores a cumplir con reglas de “compartir igual” (*share alike*), lo que la industria tecnológica comercial busca evitar (Salvagno et al., 2023).

Frente al uso no autorizado de obras protegidas, la industria ha optado por una gran falta de transparencia. Inicialmente, existía cierta apertura sobre los datos utilizados, pero en los últimos años los

desarrolladores han reducido las explicaciones detalladas a descripciones muy breves. Las grandes empresas justifican este secretismo para proteger su posición competitiva y evitar riesgos de seguridad, argumentando que la transparencia facilitaría la copia de sus innovaciones (Ihalainen, 2018).

Desde la perspectiva de las políticas públicas, Buick (2025) afirma que la falta de transparencia dificulta la detección de sesgos o discriminación en los datos. Además, se considera que el verdadero motivo del secretismo es reducir el riesgo legal por posibles infracciones de derechos de autor. Algunos informes señalan que muchas empresas no conservan registros internos de sus datos de entrenamiento por temor a que se utilicen como prueba en litigios.

Para contrarrestar esta asimetría, la academia y las coaliciones de autores han propuesto varios modelos de transparencia. El enfoque de “acceso total” (*full access*), que exige hacer públicos los conjuntos de datos, es inviable tanto por su tamaño monumental como porque publicarlos supondría una infracción masiva de derechos de autor y de las normativas de protección de datos personales.

Alternativas como el “acceso restringido” (*gated access*), sujeto a verificación de credenciales, presentan grandes retos de infraestructura. Por ello, el consenso regulatorio favorece la exigencia de “resúmenes de datos”, como las etiquetas de nutrición de datos (*dataset nutrition labels*). Sin embargo, si estos resúmenes no incluyen metadatos específicos de cada obra, la capacidad de los titulares para verificar el uso de sus obras sigue siendo casi nula. Esto demuestra que la transparencia, por sí sola, no resuelve el problema si las estructuras jurídicas no se adaptan adecuadamente.

El debate internacional sobre la legalidad del uso de datos protegidos para entrenar modelos ha generado una gran fragmentación entre jurisdicciones. Esto evidencia que el impacto de cualquier requisito de transparencia depende de las leyes de derechos de autor de cada país. Autores como Nieva Fenoll (2018) y Boden (2017) coinciden en que entrenar un modelo generativo implica reproducir los datos al menos una vez de forma electrónica.

En varios sistemas legales, esta reproducción técnica, aunque sea temporal o incidental, se considera una infracción del derecho exclusivo de reproducción (Spindler, 2019). Obtener o licenciar los

derechos de millones de obras para entrenar modelos sería económicamente inviable, ya que los costos de identificación y negociación con cada autor serían demasiado altos. Por ello, la industria de la inteligencia artificial depende de las excepciones y limitaciones al derecho de autor para operar (Spindler, 2019).

A nivel global, existen tres enfoques principales sobre este tema. En primer lugar, la Unión Europea cuenta con un sistema de excepciones limitadas. La Directiva 2019/790/UE es la norma más relevante, ya que sus artículos 3 y 4 permiten la minería de textos y datos. El artículo 4 es especialmente relevante para la inteligencia artificial generativa comercial, pues permite la minería de textos y datos para cualquier fin, siempre que los titulares de derechos no hayan “reservado expresamente” el uso de sus obras mediante el mecanismo de *opt-out*. Algunos autores, como Ziaja (2024) y Margoni y Kretschmer (2022), han criticado este mecanismo para usos comerciales. Buick (2025) indica que el *opt-out* no es efectivo por dos razones principales: no existe un estándar tecnológico universal para expresar esta reserva de forma legible por máquina y los titulares no pueden saber si sus obras fueron usadas, lo que les impide comprobar si su reserva fue respetada.

En contraste, Estados Unidos aplica la doctrina del *fair use* (uso legítimo), un sistema flexible evaluado caso por caso según cuatro factores legales. Los tribunales estadounidenses han destacado el “propósito y carácter del uso”, desarrollando el concepto de uso “transformativo”. Casos como *Authors Guild v HathiTrust* (2014) y *Authors Guild v Google* (2015) han establecido que el análisis computacional no expresivo tiene un propósito distinto y modifica el mensaje original, por lo que se considera un uso legal incluso si es comercial.

En el sistema de *fair use* no existe el derecho al *opt-out*, por lo que un requisito de transparencia en Estados Unidos tendría un efecto legal muy diferente al de Europa, ya que no les otorgaría a los titulares un poder de veto automático. La excepción amplia y flexible que permite la explotación de obras para “fines de no disfrute” (*non-enjoyment purposes*), incluida la minería de datos comercial, carece de cualquier mecanismo de *opt-out* para los creadores.

Singapur ha adoptado una postura similar, legalizando la copia

para el análisis computacional de datos. En cambio, países con listas cerradas de excepciones, como el Reino Unido, solo permiten la minería de textos y datos de obras protegidas para investigaciones estrictamente no comerciales, lo que genera barreras significativas para los *startups* locales de inteligencia artificial (Roberts et al., 2023).

Coincidimos con Odeh (2025) en que estas diferencias pueden entrar pronto en conflicto con el derecho internacional, especialmente con el test de las tres etapas del Convenio de Berna (*three-step test*). Este estándar internacional establece que las excepciones solo se aplican en casos especiales, siempre que no afecten la explotación normal de la obra ni perjudiquen injustamente los intereses legítimos del autor.

Dado que los modelos generativos pueden crear contenido que sustituye económicamente a los materiales originales, se argumenta que las excepciones comerciales amplias para el entrenamiento de IA, como el *fair use* o la excepción japonesa, podrían infringir la tercera etapa del test de Berna. Esta tensión demuestra que los requisitos de transparencia en los datos, aunque necesarios, enfrentan una situación legal inestable que requiere una revisión en todos los foros globales.

2.2 La EU AI Act y los límites de la transparencia

La EU AI Act es el resultado legislativo de la tensión entre tecnología y derechos de autor en Europa. El reglamento demuestra que las reglas de transparencia, sin reformas sustanciales en propiedad intelectual, no resuelven los problemas de fondo. Según Buick (2025), las obligaciones sobre derechos de autor no figuraban en la propuesta original de la Comisión Europea y se incorporaron tras el lanzamiento de ChatGPT a finales de 2022, debido a la presión del sector cultural y las industrias creativas. Por ello, el artículo 53 de la EU AI Act establece un marco regulatorio con dos pilares principales.

El artículo 53, apartado 1, letra (c) exige que los proveedores de modelos de inteligencia artificial de propósito general adopten políticas para cumplir la ley de derechos de autor de la Unión Europea. Deben identificar y respetar activamente, mediante tecnologías avanzadas, las reservas de derechos (*opt-outs*) según el artículo 4,

apartado 3 de la Directiva 2019/790/UE. El apartado (d) también obliga a crear y publicar un resumen suficientemente detallado del contenido utilizado para entrenar el modelo.

Sin embargo, estos mandatos presentan varias fallas estructurales que limitan su efectividad. El primer problema es el concepto de “resumen suficientemente detallado”. El considerando 107 del reglamento indica que, para proteger secretos comerciales y datos confidenciales, los resúmenes deben ser generales y no exhaustivos, mencionando solo bases de datos principales y no obras específicas. Por lo tanto, los autores no podrán saber si sus obras fueron utilizadas por el modelo, lo que les impide ejercer el *opt-out*. Además, el artículo 4(2) de la Directiva 2019/790/UE exige eliminar las copias creadas para minería de textos y datos tras su uso, lo que imposibilita revisar los datos originales posteriormente.

El segundo gran límite es la extraterritorialidad. La mayoría de los desarrollos de inteligencia artificial se realizan fuera de la Unión Europea. El considerando 106 establece que las reglas de transparencia y respeto del *opt-out* deben aplicarse a cualquier proveedor que desee acceder al mercado europeo, para garantizar igualdad de condiciones. Sin embargo, el derecho de autor es territorial. Si una empresa realiza *web scraping* y recopilación de datos en un país donde esto es legal —como Japón o bajo una interpretación amplia del *fair use* en Estados Unidos— no existe infracción de derechos de autor en Europa.

Si la EU AI Act prohíbe que ese modelo entre al mercado europeo, estaría imponiendo, como sostienen Liesenfeld y Dingemanse (2024), un efecto extraterritorial radical “por la puerta trasera” de un considerando no vinculante, lo que es muy inusual en la teoría del derecho de autor.

Por último, las barreras logísticas y económicas limitan la efectividad de la norma para los creadores individuales. Aunque se propone la gestión colectiva de derechos para facilitar licencias masivas y gestionar los *opt-outs*, la cantidad de obras en un LLM supera ampliamente los repertorios gestionados anteriormente, como en la música. Administrar miles de millones de obras requeriría infraestructuras enormes financiadas por los Estados, que actualmente no existen. Por ello, es probable que las grandes empresas tecnológi-

cas eviten la incertidumbre mediante acuerdos de licencias multimillonarios con grandes grupos editoriales que controlan extensos archivos de contenido. Así, los proveedores de inteligencia artificial cumplirán formalmente con la EU AI Act, pero los autores originales seguirán excluidos del sistema económico, lo que demuestra que la transparencia, aunque necesaria, no es suficiente para equilibrar la innovación y la remuneración justa.

3. Análisis de la petición de decisión prejudicial en el Asunto C-250/25 (*Like Company c. Google*)

3.1 Contexto y alcances de las cuestiones sobre la comunicación al público y la generación de respuestas

El asunto C-250/25 (*Like Company contra Google Ireland Limited*) presenta al Tribunal de Justicia de la Unión Europea un desafío legal clave de la era digital: la relación entre los modelos de inteligencia artificial y los derechos de propiedad intelectual. La cuestión principal es si mostrar un texto parcialmente idéntico al de los editores de prensa, generado por un chatbot basado en un LLM, constituye un acto de comunicación al público.

Para comprender la relevancia de este tema, es importante señalar que la generación de respuestas por un LLM cuestiona los fundamentos tradicionales del derecho de autor. Como indica Lemley (2024), la inteligencia artificial generativa “pone patas arriba el derecho de autor” [*turns copyright upside down*] (p. 190), porque elimina las distinciones entre uso transformador, copia literal e imitación de estilos o ideas.

El tribunal húngaro que remitió la cuestión señala que las respuestas de estos chatbots se generan mediante un proceso probabilístico, en el que el sistema predice la siguiente palabra según los patrones aprendidos. Por tanto, el modelo no recupera archivos guardados para mostrarlos directamente al público, sino que “alucina” o construye respuestas de forma probabilística (Achiam et al., 2023; Brown et al., 2020; Yan et al., 2024).

El debate sobre si este mecanismo infringe el derecho de comunicación al público es global. A nivel internacional, casos como *The New York Times Company v. OpenAI LP* (2023) argumentan que la

capacidad de los modelos para reproducir casi literalmente partes de artículos protegidos puede suponer una apropiación del mercado de los editores. La doctrina establece que, al evaluar estos resultados, debe analizarse si el uso compite directamente con la obra original o la reemplaza en el mercado (Samuelson, 2023).

Según la Directiva 2001/29/CE, la comunicación al público exige un acto de transmisión y la existencia de un público nuevo. No obstante, aplicar este criterio a una interfaz conversacional basada en un LLM, donde el texto se genera dinámicamente para un solo usuario, requiere una interpretación cuidadosa que equilibre la protección de la innovación tecnológica y la inversión de los editores de prensa.

3.2 La fase de entrenamiento de modelos LLM como presunto acto de reproducción

La segunda cuestión prejudicial analiza si el entrenamiento de un chatbot basado en un LLM, que identifica y empareja patrones, puede considerarse un acto de reproducción. El aprendizaje profundo requiere grandes volúmenes de datos, obtenidos frecuentemente mediante el raspado web de repositorios como Common Crawl. Estos datos permiten ajustar miles de millones de parámetros matemáticos (Goodfellow et al., 2016). Durante el proceso, obras protegidas por derechos de autor se copian temporalmente en la memoria RAM y se procesan para extraer reglas sintácticas, semánticas y lógicas.

El debate se centra en si estas copias técnicas incidentales constituyen reproducciones infractoras según el artículo 2 de la Directiva InfoSoc. Autores como Quang (2021) y Chen (2023) argumentan que el uso de obras protegidas para entrenar inteligencia artificial infringe derechos exclusivos, ya que el valor comercial del modelo depende de la expresividad del corpus de entrenamiento. Sin embargo, es preciso diferenciar entre la fase de ingesta (*input*) y la de generación (*output*). Como señala Sag (2023), el derecho de autor no debería otorgarles a los titulares un monopolio sobre la extracción de hechos, ideas o correlaciones estadísticas no expresivas presentes en sus obras. Por ello, la copia no consuntiva propia del aprendizaje automático debería considerarse, en principio, un uso permitido o justificado.

El debate técnico y dogmático se complica con el fenómeno de la

“memorización algorítmica”. Aunque el objetivo del entrenamiento es la generalización estadística, los modelos de lenguaje pueden sobreajustarse cuando se entrenan con datos que contienen secuencias muy redundantes, duplicadas o con baja entropía. En estas condiciones, la función de optimización ajusta los parámetros para que las probabilidades condicionales converjan en trayectorias casi deterministas, lo que permite que técnicas de extracción o ciertos *prompts* induzcan al modelo a reconstruir, de forma idéntica o casi idéntica, fragmentos extensos del material original (Zhou et al., 2024).

Este fenómeno es especialmente frecuente en textos altamente estructurados o en código fuente informático (Emanuilov y Margoni, 2024). Desde la perspectiva del derecho de autor, esta retención latente en los espacios vectoriales desafía la distinción tradicional entre la extracción computacional de correlaciones abstractas y la fijación duradera de una expresión protegida, lo que complica el análisis sobre si la ponderación paramétrica puede constituir una forma indirecta de reproducción no autorizada.

Si el modelo retiene la expresión original en sus parámetros, la distinción entre extraer ideas y copiar expresiones se vuelve menos clara. Por ello, la transparencia sobre los datos de entrenamiento es clave para cualquier análisis jurídico. Sin embargo, la falta de información por parte de las empresas sobre los conjuntos de datos dificulta que los titulares de derechos prueben qué obras se reprodujeron durante el entrenamiento (Jernite, 2023). Así, la decisión del Tribunal de Justicia de la Unión Europea (TJUE) sobre este tema redefinirá la legalidad del proceso industrial detrás de la inteligencia artificial en Europa.

3.3 La excepción de minería de textos y datos y la reproducción impulsada por el usuario

La tercera y la cuarta cuestión prejudicial vinculan el análisis técnico con las excepciones legales y la responsabilidad del usuario final. Si el entrenamiento de modelos se considera reproducción, el tribunal consulta si este acto está amparado por la excepción de minería de textos y datos del artículo 4 de la Directiva 2019/790/UE. La minería de datos representa un cambio relevante: el valor proviene del análisis

sis computacional, no del consumo humano de la obra. La doctrina europea ha analizado cómo esta excepción busca equilibrar la competitividad tecnológica y la protección de los derechos de autor, permitiendo la minería con fines comerciales siempre que el titular no haya reservado sus derechos mediante un *opt-out* legible por máquina (Ducato y Strowel, 2021; Margoni y Kretschmer, 2022; Ziaja, 2024).

Para determinar la sostenibilidad de este marco a largo plazo, es necesario compararlo con otros sistemas legales. Por ejemplo, Japón ha adoptado excepciones de derechos de autor muy flexibles para fines de “no disfrute”, lo que permite el análisis computacional masivo, incluso con fines comerciales, ya que se considera que la inteligencia artificial no “percibe” la obra de forma estética o intelectual (Dermawan, 2024; Ueno, 2021).

En Estados Unidos, la doctrina del *fair use* ha respaldado grandes proyectos de digitalización y ha establecido el concepto “aprendizaje justo” (*fair learning*), donde el uso de obras para entrenar algoritmos se considera muy transformador y legal (*Authors Guild v. Google, Inc.*, 2015; *Authors Guild v. HathiTrust*, 2014; Lemley y Casey, 2021).

La cuarta cuestión plantea la complejidad de la interacción humana: ¿es responsable el proveedor del chatbot si el usuario introduce una instrucción (*prompt*) que provoca que el sistema reproduzca contenido protegido?

Esto exige analizar la responsabilidad directa frente a la indirecta o contributiva. El proveedor de inteligencia artificial diseña el modelo, pero el usuario genera la respuesta específica. Si el usuario emplea el modelo para eludir muros de pago o copiar artículos de prensa, las reglas tradicionales de exención de responsabilidad para intermediarios tecnológicos se ponen a prueba (Senftleben, 2023). Por tanto, resolver este problema implica determinar si los modelos de inteligencia artificial son herramientas neutras o si facilitan activamente la infracción cuando reciben instrucciones extractivas.

4. Crítica a la admisibilidad y la precisión técnica de la remisión

4.1 Problemas de admisibilidad y la naturaleza hipotética del litigio

El análisis procesal de la petición en el caso C-250/25 revela problemas significativos que cuestionan su admisibilidad ante el TJUE. Según la European Copyright Society (2026), las cuestiones prejudiciales deben resolver disputas concretas, no debates académicos o hipótesis tecnológicas. En este caso, la falta de precisión sobre la jurisdicción aplicable a los actos de entrenamiento, sumada a la probabilidad de que el modelo de Google no se haya entrenado en Hungría, hace que aplicar la ley húngara y la normativa europea de derechos de autor resulte dudoso y que el litigio sea especulativo.

Consideramos que esta consulta judicial refleja el esfuerzo apresurado y fragmentado de Europa por establecer reglas claras para la inteligencia artificial. Utilizar un litigio limitado —centrado en el derecho de los editores de prensa— para determinar el estatus legal de los modelos fundacionales puede generar decisiones legales fuera de contexto. Además, la entrada en vigor de la EU AI Act ha hecho que el marco normativo europeo sea más complejo y esté más vinculado al derecho de autor.

Esta misma norma exige transparencia y respeto a la normativa de derechos de autor, incluyendo el cumplimiento de las reservas de derechos (*opt-outs*) del artículo 4 de la Directiva 2019/790/UE para los proveedores de modelos de inteligencia artificial de propósito general, sin importar dónde se haya realizado el entrenamiento, siempre que el modelo se comercialice en Europa (Espinoza, 2024; Liesenfeld y Dingemanse, 2024; Peukert, 2024).

Otro aspecto relevante es que, para evaluar la legalidad de las operaciones de una gran empresa tecnológica, se deben considerar tanto las directivas tradicionales como el nuevo marco regulatorio. Tomar decisiones basadas en hechos poco claros puede paralizar la innovación, convirtiendo el derecho de autor en una barrera en lugar de un apoyo, un equilibrio históricamente difícil de alcanzar en Europa (Rosati, 2024).

El principal problema técnico de la resolución judicial es que agrupa tecnologías distintas bajo un solo término. La resolución confunde los conceptos de “chatbot”, “modelo de lenguaje grande” y

“motor de búsqueda”. Desde la ingeniería de *software* y la regulación, estos elementos son partes separadas de la arquitectura tecnológica y tienen implicaciones legales diferentes (Anthropic PBC, 2023). Un LLM es el modelo base entrenado con grandes volúmenes de datos; el chatbot (como Gemini o ChatGPT) es la interfaz de usuario para el diálogo; el motor de búsqueda es el servicio que recupera información en tiempo real.

La falta de detalle al describir el “proceso de entrenamiento” y los datos utilizados distorsiona el análisis legal. La gobernanza de los datos es clave; se deben documentar adecuadamente los conjuntos de datos, sus fuentes y posibles sesgos antes de evaluar la salida de un modelo (Gebbru et al., 2021; Rosati, 2019).

Cuando un tribunal asume que un chatbot se entrena “mediante la búsqueda de cadenas”, evidencia una falta de comprensión de las matemáticas de los espacios latentes y los vectores que permiten el aprendizaje profundo. Además, la tendencia de la industria hacia modelos cerrados u opacos incrementa esta confusión en los tribunales. Algunos líderes del sector han reconocido que su estrategia inicial de apertura fue “equivocada” desde el punto de vista competitivo (Vincent, 2023), y juristas como Lessig (2024) advierten sobre los riesgos cívicos de mantener monopolios cerrados sobre los modelos fundacionales.

Para resolver esta confusión, se requiere un nivel de transparencia en los datos de entrenamiento que permita una auditoría técnica y legal efectiva (Li, 2023). También es necesario promover conjuntos de datos gestionados como bienes comunes y entrenar modelos abiertos únicamente con imágenes y textos libres de derechos o bajo licencias Creative Commons, como en el proyecto CommonCanvas. Estas medidas permitirán avanzar tecnológicamente sin infringir derechos de forma sistemática (Gokaslan et al., 2024; Lessig, 2004; Purtova y Van Maanen, 2024; Szkalej y Senftleben, 2023). Si no se define claramente qué parte del sistema técnico es responsable de cada acción sobre los datos, no se puede evaluar la responsabilidad de manera objetiva.

4.2 La distinción crítica entre el entrenamiento de modelos y la generación aumentada por recuperación

El último aspecto clave para comprender este escenario técnico-jurídico es diferenciar el entrenamiento de pesos de un LLM de la técnica de generación aumentada por recuperación (*retrieval-augmented generation*, RAG). En la remisión prejudicial del asunto C-250/25, se observa que el sistema evaluado utiliza información actualizada dinámicamente mediante consultas a fuentes externas al responder. Desde la perspectiva de la arquitectura, RAG no modifica ni reentrena la matriz de parámetros estáticos del modelo fundacional. En cambio, funciona en dos etapas: primero, un módulo de recuperación (*retriever*) realiza una búsqueda, generalmente por similitud de cosenos en una base de datos vectorial o mediante motores de indexación web en tiempo real, para extraer fragmentos de documentos externos relevantes para la consulta del usuario; segundo, el sistema inserta estos fragmentos en la “ventana de contexto” (*context window*) del *prompt* enviado al modelo generativo (*generator*). Así, el LLM emplea su capacidad lingüística solo para sintetizar, formar y generar una respuesta basada en los textos recuperados en ese momento.

Calificar el proceso de RAG como “entrenamiento” es incorrecto y dificulta la aplicación de las excepciones de minería de textos y datos, como señaló el tribunal húngaro. El uso de datos editoriales en RAG se asemeja más al funcionamiento avanzado de un motor de búsqueda o un agregador de noticias. Si un sistema utiliza contenido periodístico protegido sin licencia en su memoria temporal para crear un resumen, la calidad del resultado depende directamente de la calidad del contenido original, reflejando el principio informático “basura entra, basura sale” (*garbage in, garbage out*), en este caso, aplicado al valor económico de la información periodística confiable (De la Durantaye, 2023).

Las repercusiones económicas de esta distinción son significativas. El auge de la inteligencia artificial generativa ha impulsado análisis económicos sobre la relación entre tecnología y derecho de autor, anticipando disrupciones importantes en los modelos de negocio creativos (Bloomberg Intelligence, 2024; Lu y Zhou, 2021).

Si los chatbots con RAG pueden sintetizar artículos de prensa de

forma que desincentiven al usuario de visitar la fuente original, el modelo de ingresos de los editores basado en publicidad y suscripciones se ve amenazado. Ante este riesgo y la vulnerabilidad jurídica de sus sistemas, empresas líderes han comenzado a negociar acuerdos de licencia multimillonarios con los principales conglomerados de medios para legitimar el uso de sus archivos periodísticos y el suministro de datos en tiempo real (Lopatto, 2024).

En última instancia, la regulación y la jurisprudencia deben adoptar un enfoque pragmático que establezca prioridades claras, diferenciando entre la libertad técnica para entrenar algoritmos matemáticos y la obligación legal de licenciar contenido cuando este se utiliza y compete comercialmente en el mercado de la información (Lim, 2023; Senftleben, 2023).

5. El derecho de los editores de prensa y las excepciones de minería de datos

5.1 Delimitación del sujeto protegido en el marco del artículo 15 de la Directiva 2019/790/UE

Para analizar la controversia del caso C-250/25 (*Like Company c. Google*), es esencial definir la naturaleza jurídica del derecho en cuestión. El tribunal fundamenta sus preguntas en el derecho de los editores de prensa, conocido como “derecho conexo” o Press Publishers’ Right (PPR), mencionado en el artículo 15 de la Directiva 2019/790/UE. Esta distinción es relevante porque el derecho conexo protege la inversión organizativa y financiera para crear y mantener publicaciones de prensa, no la originalidad ni la personalidad del autor, que son el núcleo del derecho de autor clásico.

La diferencia entre los derechos de autor originales y los derechos conexos tiene una base sólida en el derecho internacional de propiedad intelectual. Documentos clave como el Convenio de Berna de 1886 (arts. 5, 8, 9, 12) y el Acuerdo sobre los ADPIC (1994, art. 12) establecen una protección mínima basada en la creación intelectual, mientras que la protección de inversiones industriales queda en manos de normas paralelas o específicas.

Las declaraciones del Tratado de la OMPI sobre Derecho de Autor refuerzan esta separación y aseguran que los derechos conexos

no reduzcan la protección del derecho de autor sobre las obras originales (World Intellectual Property Organization, 1996). En el caso de publicaciones periodísticas, cada artículo está protegido por el derecho de autor de su autor, y el artículo 15 de la directiva europea añade una protección extra para el editor sobre la colección u organización de esos artículos.

El avance de la inteligencia artificial generativa ha desafiado este sistema dual. Desde la perspectiva económica del derecho, el artículo 15 pretende evitar que los agregadores digitales obtengan beneficios sin aportar valor, una lógica que ahora se extiende a los desarrolladores de LLM (Lu y Zhou, 2021).

Si la inteligencia artificial utiliza y reproduce contenido periodístico sin compensación, existe el riesgo de colapso del mercado informativo, como señala la teoría de la asimetría de la información. Según el modelo de Akerlof (1970) sobre el mercado de los “limones”, desarrollado por Kerton y Bodell (1995), si los editores de calidad no reciben ingresos porque la inteligencia artificial genera resúmenes gratuitos, podrían abandonar el mercado. Esto dejaría a los consumidores expuestos a contenido sintético de baja calidad, desinformación o “limones” periodísticos.

Esta dinámica económica explica la intensidad de litigios transatlánticos similares, como *The New York Times Company v. OpenAI LP* (2023). En este proceso estadounidense, el demandante invoca tanto el interés organizativo como la titularidad de los derechos de autor de sus periodistas y demanda por la infracción de la obra original, acusando a la inteligencia artificial de usurpar el mercado mediante la memorización de artículos.

En contraste, en el caso *Like Company*, la remisión se limita al PPR. La doctrina advierte que abordar la crisis de la inteligencia artificial generativa solo desde el derecho de los editores es una estrategia jurídicamente asimétrica, ya que deja desprotegidos a los creadores individuales y debilita el enfoque de utilidad social del derecho de autor (Gervais, 2022).

Históricamente, las tensiones de uso masivo se han gestionado mediante sociedades de gestión colectiva (Ficsor, 2002), un mecanismo potencialmente más adecuado para administrar licencias de entrenamiento de inteligencia artificial que la invocación fragmen-

tada de derechos conexos. Por tanto, el TJUE debe interpretar el artículo 15 de manera restrictiva, limitándose a proteger la inversión del editor frente a los prestadores de servicios de la sociedad de la información, sin extrapolar este derecho conexo para establecer un precedente universal sobre la licitud de la inteligencia artificial respecto al derecho de autor original.

5.2 Los límites del “uso en línea” y las excepciones al derecho de los editores

La exclusividad de explotación económica que el artículo 15 de la Directiva 2019/790/UE les concede a los editores de prensa está claramente limitada. Solo se aplica a determinados actos y contempla excepciones destinadas a salvaguardar el libre flujo de información y la diversidad democrática.

En concreto, la norma protege únicamente frente al “uso en línea” de publicaciones de prensa por proveedores de servicios de la sociedad de la información. Este concepto fue diseñado para abarcar motores de búsqueda y agregadores de noticias, que reutilizan titulares y resúmenes periodísticos. Por tanto, equiparar la interfaz de un chatbot basado en un LLM con la de un agregador de noticias no resulta adecuado desde una perspectiva técnica ni legal.

En el caso de sistemas como ChatGPT o Gemini (analizado en las cuestiones prejudiciales 1 y 4), destaca la excepción que excluye de protección el uso de “palabras sueltas o de extractos muy breves de publicaciones de prensa”. Por ello, determinar qué constituye un “extracto muy breve” en un modelo algorítmico representa un desafío relevante.

La informática ha demostrado que, en determinadas circunstancias, las redes neuronales pueden memorizar y reproducir fragmentos exactos de sus datos de entrenamiento (Zhou et al., 2024). Si un LLM, tras recibir un *prompt*, reproduce un párrafo completo de una noticia, excede la excepción y constituye una infracción de los derechos de reproducción y comunicación pública.

Sin embargo, el tratamiento de estas excepciones varía significativamente a nivel internacional, lo que sitúa a la Unión Europea en una posición compleja tanto en términos competitivos como éticos.

En Estados Unidos, la reproducción no autorizada de obras protegidas con fines tecnológicos suele evaluarse bajo la doctrina del *fair use*, establecida en casos como *Campbell v. Acuff-Rose Music, Inc.* (1994).

Esta doctrina ha originado el concepto de aprendizaje justo, que considera transformador el uso de materiales protegidos para entrenar sistemas computacionales, siempre que no se utilice la expresión artística original, sino únicamente sus patrones estadísticos (Lemley y Casey, 2021).

De manera similar, países como Japón han adoptado excepciones amplias para fines de “no disfrute”, permitiendo el escaneo y análisis computacional independientemente de si existen fines comerciales, y priorizando el desarrollo de la inteligencia artificial sobre el control absoluto del titular (Dermawan, 2024; Ueno, 2021).

En Europa, la rigidez normativa y la estricta limitación a “extractos muy breves” podrían privar a los sistemas de inteligencia artificial de acceso a datos de calidad. Como indica la literatura técnica y jurídica, la calidad de un modelo algorítmico depende directamente de la calidad de sus datos: “basura entra, basura sale” (*garbage in, garbage out*) (De la Durantaye, 2023).

Si el TJUE interpreta el límite del uso en línea de forma tan amplia que impida el acceso a publicaciones de prensa mediante técnicas como la RAG, la calidad de las respuestas de las inteligencias artificiales europeas se verá afectada.

Además, una restricción tan estricta plantea un problema relevante sobre el reparto de la remuneración entre los autores, ya que el artículo 15 exige que los editores compartan los ingresos con los periodistas (Senftleben, 2023). Si las empresas tecnológicas optan por bloquear el contenido europeo en lugar de licenciarlo para evitar riesgos legales (Korinek, 2023; Lim, 2023), la norma diseñada para proteger a la prensa podría acabar perjudicando tanto a editores como a autores en el entorno digital.

5.3 La aplicación de la excepción de minería de textos y datos (artículo 4) a la inteligencia artificial

Si el uso masivo de documentos para entrenar modelos fundacionales se considera un acto de reproducción relevante según la Directi-

va 2019/790/UE, la legalidad de la inteligencia artificial en Europa dependerá de las excepciones para la minería de textos y datos, en particular del artículo 4. Esta norma permite copiar y extraer obras a las que se tiene acceso legal para minería de textos y datos, incluso con fines comerciales, salvo que los titulares de derechos hayan reservado su uso de forma clara, por ejemplo, mediante sistemas legibles por máquina (*opt-out*). La excepción se diseñó para facilitar la investigación y el análisis de big data, aunque varios autores la consideran demasiado limitada frente a leyes internacionales más flexibles (Ducato y Strowel, 2021; Margoni y Kretschmer, 2022; Rosati, 2019).

La aparición de los LLM ha situado al artículo 4 en el centro del debate regulatorio. Esta relevancia se refleja en la EU AI Act, donde el legislador europeo, al buscar armonizar distintas normas, exige (art. 53, apartado 1, letra c) que los proveedores de modelos de inteligencia artificial de propósito general cuenten con políticas para respetar el derecho de autor de la Unión y, en particular, identifiquen y cumplan con las reservas de derechos según el artículo 4, apartado 3 de la Directiva 2019/790/UE sobre los Derechos de Autor en el Mercado Único Digital (Liesenfeld y Dingemanse, 2024; Peukert, 2024).

Sin embargo, la aplicación práctica del *opt-out* presenta importantes desafíos. Los expertos debaten si la reserva de derechos para minería de textos y datos constituye un veto legítimo para proteger la soberanía del creador o si, por el contrario, representa una barrera que puede limitar la innovación europea al fragmentar los conjuntos de datos de entrenamiento (Ziaja, 2024). Las guías para establecer buenas prácticas de *opt-out* aún se encuentran en una fase inicial (Keller y Warsó, 2023). Al mismo tiempo, la demanda de transparencia sobre los datos de entrenamiento aumenta tanto a nivel social como institucional. Nuevas regulaciones, como la California AI Transparency Act (S.B. 942, 2024) en Estados Unidos y los requisitos de resúmenes detallados de la EU AI Act buscan clarificar qué obras utilizan los algoritmos de las grandes empresas tecnológicas (Jernite, 2023; Li, 2023).

En este contexto de tensión entre el *opt-out* y la alta demanda de datos por parte de los algoritmos, evidenciada por la dependencia de la industria en repositorios masivos como Common Crawl, sur-

gen modelos alternativos de gobernanza. Por un lado, se investiga el uso de datos sintéticos generados por la propia inteligencia artificial, aunque esto podría generar problemas futuros (Goyal y Mahmoud, 2024). Por otro, cobra relevancia la gestión de conjuntos de datos basados en bienes comunes digitales (Purtova y Van Maanen, 2024).

Así, algunos proyectos que entrenan modelos abiertos únicamente con imágenes libres o bajo licencias Creative Commons (Gokaslan et al., 2024; Lessig, 2004) representan una alternativa ética. Sin embargo, aplicar cláusulas como “compartir igual” a los modelos entrenados y a sus resultados sigue siendo un desafío legal complejo y poco explorado (Szkalej y Senftleben, 2024). En conclusión, consideramos necesario que el TJUE evalúe cuidadosamente si la fase de entrenamiento en el caso *Like Company* cumple con el artículo 4. Una decisión restrictiva sin tener en cuenta el test de las tres etapas que define las excepciones (Rosati, 2024) podría provocar que el desarrollo de la inteligencia artificial se traslade fuera de la Unión Europea.

5.4 A modo de discusión

El derecho de la propiedad intelectual se enfrenta de lleno a la revolución algorítmica, y el litigio prejudicial C-250/25 es solo un ejemplo dentro de una crisis global más amplia. Como señala Lemley (2024), la inteligencia artificial generativa ha cambiado por completo los fundamentos del derecho de autor.

Anteriormente, la ley se enfocaba en la copia como un fin, destinada al disfrute estético o intelectual de una obra pirateada. Actualmente, la copia se utiliza como una herramienta mecánica e industrial para identificar correlaciones matemáticas y desarrollar modelos probabilísticos que imitan la cognición humana. Este cambio desafía al sistema legal a distinguir entre la protección de la creatividad y la permisividad hacia la monopolización de datos, fundamentales para el progreso tecnológico.

El caso analizado demuestra cómo la industria tecnológica ha pasado de un paradigma de ciencia abierta a un hermetismo corporativo, justificado por la búsqueda de ventaja competitiva y seguridad. Líderes del sector reconocen este cambio como un error significativo respecto a sus principios fundacionales de apertura (Vincent, 2023).

Esta concentración de poder en modelos cerrados (*closed-source*) es vista con alarma por la doctrina legal, que sostiene que no todos los modelos de inteligencia artificial fundacionales deberían operar bajo el control exclusivo de un puñado de corporaciones, pues se corre el riesgo de privatizar la infraestructura del conocimiento humano (Lessig, 2024).

En este contexto, en lugar de fomentar una inteligencia artificial abierta, los conflictos sobre derechos de autor están generando una nueva “economía de licencias”. Por temor a litigios costosos, las grandes empresas han firmado acuerdos millonarios con editoriales para obtener acceso exclusivo a datos en tiempo real y derechos de entrenamiento. Esto excluye a *startups*, universidades y desarrolladores de código abierto que carecen de los recursos para negociar acuerdos similares (Bloomberg Intelligence, 2024; Lopatto, 2024).

Esta situación amplía una brecha socioeconómica preexistente: la pérdida de empleo en la clase creativa. La inteligencia artificial generativa, además de competir con los grandes medios, impacta significativamente a *freelancers*, ilustradores, redactores y programadores, reduciendo la demanda de trabajo humano en plataformas digitales (Demirci et al., 2025).

La frustración de la comunidad creativa es evidente a nivel global. Instituciones representativas han solicitado reiteradamente ante foros gubernamentales la creación de mecanismos que aseguren el consentimiento informado, el reconocimiento y una remuneración justa por el uso de su trabajo en el desarrollo de estos modelos (Anthropic PBC, 2023; Thambaiya et al., 2025).

El derecho de los editores de prensa, examinado en el caso *Like Company*, no aborda las vulnerabilidades existenciales de los creadores individuales; únicamente redistribuye rentas entre grandes empresas mediáticas y tecnológicas.

Desde una perspectiva geopolítica, los Estados siguen estos conflictos legales con mucho cuidado, ya que ven a la inteligencia artificial como clave para la seguridad nacional y la economía. Las grandes potencias están creando estrategias nacionales para liderar en inteligencia artificial (Csernatoní et al., 2025), emitiendo órdenes ejecutivas para asegurar un desarrollo tecnológico seguro y competitivo (Sabouri y Mehrdel, 2024) y participando en foros in-

ternacionales para acordar reglas que equilibren la protección de la propiedad intelectual con la necesidad de innovación global (Bode y Qiao-Franco, 2024). No en vano, si el TJUE interpreta de forma desproporcionada el derecho de los editores de prensa y equipara el entrenamiento de modelos neuronales o la técnica RAG con la distribución de copias piratas, Europa podría quedar aislada de la frontera tecnológica en contra del objetivo de las directivas que buscan fortalecer el Mercado Único Digital.

En resumen, el análisis del caso *Like Company c. Google* revela la fragilidad de las categorías jurídicas decimonónicas ante infraestructuras informáticas probabilísticas de propósito general. La insistencia de los tribunales nacionales en adaptar forzosamente tecnologías complejas y diversas a los limitados marcos de directivas redactadas en la era pregenerativa crea inseguridad jurídica. El principal desafío del derecho de autor será entonces el de rediseñar los incentivos económicos para que la creación biológica —como el talento humano, el periodismo de investigación y la expresión artística— siga siendo financieramente viable en un ecosistema digital saturado de elocuencia sintética.

6. Conclusiones

El caso C-250/25 (*Like Company contra Google*) evidencia los riesgos de la falta de claridad judicial. La remisión confunde las etapas de entrenamiento algorítmico, la estructura de los LLM y la recuperación de información con técnicas como la RAG. Modificar el alcance de los derechos exclusivos con base en hechos poco claros puede obstaculizar el desarrollo de la inteligencia artificial en Europa. Además, si los tribunales no distinguen entre el uso estadístico de datos y la comunicación de contenidos, podrían emitir decisiones que perjudiquen tanto la innovación tecnológica como el mercado digital.

El derecho conexo de los editores de prensa, establecido en el artículo 15 de la Directiva 2019/790/UE, resulta insuficiente para abordar los desafíos que plantea la inteligencia artificial. Esta protección se limita a la inversión de las empresas mediáticas y excluye a periodistas y autores. Abordar los problemas del sector creativo solo desde la perspectiva de los editores favorece una economía de

licencias dominada por grandes empresas tecnológicas y medios, sin asegurar una remuneración justa para los creadores individuales.

El sistema europeo de excepciones para la minería de textos y datos, fundamental para el entrenamiento de algoritmos, presenta problemas significativos. Aunque el artículo 4 de la Directiva 2019/790/UE sobre los Derechos de Autor en el Mercado Único Digital permite el uso comercial, el mecanismo de *opt-out* resulta poco práctico debido a la falta de protocolos estandarizados y universales legibles por máquina.

El análisis muestra que las recientes exigencias de transparencia de la EU AI Act, aunque bienintencionadas, no benefician a los autores. La obligación de proporcionar un “resumen suficientemente detallado” de los datos de entrenamiento (artículo 53) no ofrece el nivel de detalle necesario para que un autor individual verifique el uso de su obra y ejerza su *opt-out*. En la práctica, estas normas aumentan las cargas regulatorias sin generar un beneficio económico directo para los creadores individuales.

Por último, la regulación europea, estricta y fragmentada, es menos flexible que enfoques como el *fair use* en Estados Unidos o las excepciones para “no disfrute” en Asia. Dado que el derecho de autor es territorial, la rigidez de las normas europeas puede llevar a los desarrolladores a entrenar sus modelos fuera de la Unión Europea, eludiendo sus protecciones y rezagando a Europa en tecnología y economía. Por ello, es urgente que el derecho de autor evolucione hacia modelos como la gestión colectiva de licencias de datos para lograr un equilibrio real entre el avance de la inteligencia artificial y el desarrollo humano.

Bibliografía

- Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I. ... y Zopg, B. (2023). *GPT-4 technical report*. arXiv. <https://arxiv.org/abs/2303.08774>
- Akerlof, G. (1970). The market for “lemons”: Quality uncertainty and the market mechanism. *The Quarterly Journal of Economics*, 84(3), 488-500.

- Anthropic PBC. (2023). *Notification of inquiry regarding artificial intelligence and copyright: Public comments of Anthropic PBC*. <https://tinyurl.com/3zvw4hnh>
- Bloomberg Intelligence. (2024). *Generative AI 2024 report*. Bloomberg.
- Bode, I. y Qiao-Franco, G. (2024). The geopolitics of AI in warfare: Contested conceptions of human control. En Paul, R. E. Carmel y Cobbe, J. (Eds.), *Handbook on public policy and artificial intelligence* (pp. 281-294). Edward Elgar Publishing.
- Boden, M. A. (2017). *Inteligencia artificial*. Turner.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., 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., Winter, C., Hesse, C. ... y Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. <https://doi.org/10.48550/arXiv.2005.14165>
- Buick, A. (2025). Copyright and AI training data—transparency to the rescue? *Journal of Intellectual Property Law and Practice*, 20(3), 182-192. <https://doi.org/10.1093/jiplp/jpae102>
- Chen, Y. (2023). The legality of artificial intelligence's unauthorised use of copyrighted materials under China and U.S. law. *IDEA: The Law Review of the Franklin Pierce Center for Intellectual Property*, 63, 241-260. https://www.kiip.re.kr/webzine/2307/file/KIIP2307_file_14.pdf
- Cotino Hueso, L. y Simón Castellano, P. (2024). *Tratado sobre el reglamento de inteligencia artificial de la Unión Europea*. Aranzadi.
- Csernatoni, R., Broeders, D., Andersen, L. H., Hoijsink, M., Bode, I., Lindsay, J. R. y Schwarz, E. (2025). Myth, power, and agency: Rethinking artificial intelligence, geopolitics and war. *Minds and Machines*, 35(3), artículo 37. <https://doi.org/10.1007/s11023-025-09741-0>
- De la Durantaye, K. (2023). *Garbage in, garbage out*. SSRN. <https://doi.org/10.2139/ssrn.4572952>
- Demirci, O., Hannane, J. y Zhu, X. (2025). Who is AI replacing? The impact of generative AI on online freelancing platforms. *Management Science*, 71(10), 8097-8108. <https://doi.org/10.1287/mnsc.2024.05420>
- Dermawan, A. (2024). Text and data mining exceptions in the development of generative AI models: What the EU member states could learn from the Japanese “nonenjoyment” purposes? *The Journal of World Intellectual Property*, 27(1), 44-68. <https://doi.org/10.1111/jwip.12285>
- Ducato, R. y Strowel, A. (2021). Ensuring text and data mining: Remaining issues with the EU copyright exceptions and possible ways out. *European Intellectual Property Review*, 43(5), 322-328.
- Emanuilov, I. y Margoni, T. (2024). *Forget me not: Memorisation in generative sequence models trained on open source licensed code*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4720990
- Espinoza, J. (16 de julio de 2024). Europe's rushed attempt to set the rules for AI. *Financial Times*. <https://tinyurl.com/4d5kbmnv>

- European Copyright Society. (4 de marzo de 2026). *Comment of the European Copyright Society on the request for preliminary ruling in Case C-250/25 (Like Company)*. <https://dx.doi.org/10.2139/ssrn.6333539>
- Ficsor, M. (2002). *Collective management of copyright and related rights* (Vol. 855). WIPO.
- Geburu, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Iii, H. D. y Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86-92. <https://dx.doi.org/10.1145/3458723>
- Gervais, D. (2022). *A social utility conception of fair use* (Vanderbilt Law Research Paper 22-35). <https://scholarship.law.vanderbilt.edu/faculty-publications/844>
- Gokaslan, A., Cooper, A. F., Collins, J., Seguin, L., Jacobson, A., Patel, M., Frankle, J., Stephenson, C. y Kuleshov, V. (2024). Commoncanvas: Open diffusion models trained on creative-commons images. En *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8250-8260). <https://doi.org/10.48550/arXiv.2310.16825>
- Goodfellow, I., Bengio, Y. y Courville, A. (2016). *Deep learning*. MIT Press.
- Goyal, M. y Mahmoud, Q. H. (2024). A systematic review of synthetic data generation techniques using generative AI. *Electronics*, 13(17), 3509. <https://doi.org/10.3390/electronics13173509>
- Ihalainen, J. (2018). Computer creativity: Artificial intelligence and copyright. *Journal of Intellectual Property Law & Practice*, 13(9), 724-728.
- Jernite, Y. (5 de diciembre de 2023). Training data transparency in AI: Tools, trends, and policy recommendations. *Hugging Face*. <https://huggingface.co/blog/yjernite/data-transparency>
- Keller, P. y Warso, Z. (2023). Defining best practices of opting out of ML training. *Open Future Brief*, 5, 1-17.
- Kerton, R. R. y Bodell, R. W. (1995). Quality, choice, and the economics of concealment: the marketing of Lemons. *Journal of Consumer Affairs*, 29(1), 1-28. <https://doi.org/10.1111/j.1745-6606.1995.tb00037.x>
- Korinek, A. (2023). Generative AI for economic research: Use cases and implications for economists. *Journal of Economic Literature*, 61(4), 1281-1317. <https://doi.org/10.1257/jel.20231736>
- Lemley, M.A. (2024). How Generative AI Turns Copyright Upside Down. *Columbia Science and Technology Law Review*, 25(190), 190-212. <https://doi.org/10.52214/stlr.v25i2.12761>
- Lemley, M. y Casey, B. (2021). Fair learning. *Texas Law Review*, 99, 743-757. <https://tinyurl.com/5652f2tu>
- Lessig, L. (29 de julio de 2024). Not all AI models should be freely available, argues a legal scholar. *The Economist*. <https://tinyurl.com/yu3u4spu>
- Li, Z. (2023). Why the European AI Act transparency obligation is insufficient. *Nature Machine Intelligence*, 5(6), 559-560. <https://www.nature.com/articles/s42256-023-00672-y>
- Liesenfeld, A. y Dingemanse, M. (2024). Rethinking open source generative AI:

- Open-washing and the EU AI Act. En *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 1774-1787). <https://doi.org/10.1145/3630106.3659005>
- Lim, D. (2023). Generative AI and copyright: Principles, priorities and practicalities. *Journal of Intellectual Property Law and Practice*, 18(12), 841-842. <https://doi.org/10.1093/jiplp/jpad081>
- Lopatto, E. (30 de agosto de 2024). OpenAI searches for an answer to its copyright problems. *The Verge*. <https://tinyurl.com/3a4hhukn>
- Lu, Y. y Zhou, Y. (2021). A review on the economics of artificial intelligence. *Journal of Economic Surveys*, 35(4), 1045-1072. <https://doi.org/10.1111/joes.12422>
- Margoni, T. y Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: Harmonisation, data ownership, and the future of technology. *GRUR International*, 71(8), 685-701. <https://doi.org/10.1093/grurint/ikac054>
- Nieva Fenoll, J. (2018). *Inteligencia artificial y proceso judicial*. Marcial Pons.
- Odeh, M. K. (2025). Copyright in the age of artificial intelligence: Navigating access to algorithmic training materials and the three-step test for text and data mining in Nigeria. *The Journal of World Intellectual Property*, 28(2), 428-470. <https://doi.org/10.1111/jwip.12342>
- Peukert, A. (2024). Copyright in the Artificial Intelligence Act – A primer. *GRUR International*, 73(6), 497-507. <https://doi.org/10.1093/grurint/ikae057>
- Purtova, N. y Van Maanen, G. (2024). Data as an economic good, data as a commons, and data governance. *Law, Innovation and Technology*, 16(1), 1-42. <https://doi.org/10.1080/17579961.2023.2265270>
- Quang, J. (2021). Does training AI violate copyright law? *Berkeley Technology Law Review*, 36, 1407.
- Roberts, H., Babuta, A., Morley, J., Thomas, C., Taddeo, M. y Floridi, L. (2023). Artificial intelligence regulation in the United Kingdom: A path to good governance and global leadership? *Internet Policy Review*, 12(2), 1-31. <https://www.econstor.eu/handle/10419/278798>
- Rosati, E. (2019). Copyright as an obstacle or an enabler? A European perspective on text and data mining and its role in the development of AI creativity. *Asia Pacific Law Review*, 27(2), 198-217. <https://doi.org/10.1080/10192557.2019.1705525>
- Rosati, E. (2024). No step-free copyright exceptions: The role of the three-step in defining permitted uses of protected content. *European Intellectual Property Review*, 46, 262.
- Sabouri, Z. y Mehrdel, S. B. (2024). The new geopolitics of artificial intelligence and the challenges of global governance. *International Journal of Political Science*, 14(3), 149-169. <https://oiccpres.com/ijps/article/view/18589>
- Sag, M. (2023). Copyright safety for generative AI. *Houston Law Review*, 61(2), 295-321.

- Salvagno, M., Taccone, F. S. y Gerli, A. G. (2023). Can artificial intelligence help for scientific writing? *Critical Care*, 27(1), artículo 75. <https://doi.org/10.1186/s13054-023-04380-2>
- Samuelson, P. (2023). Generative AI meets copyright. *Science*, 381(6654), 158-161. <https://doi.org/10.1126/science.adi0656>
- Santarcangelo, V., Lamacchia, A., Massa, E., Gianluca, S., Crisafulli, M. G. y Basile, V. (2023). Sustainability explained by ChatGPT artificial intelligence in a HITL perspective: Innovative approaches. En *SIS 2023 Statistical learning sustainability and impact evaluation: Book of the short papers* (pp. 881-886). Pearson.
- Senftleben, M. (2023). Generative AI and author remuneration. *International Review of Intellectual Property and Competition Law*, 54, 1535-1561. <https://doi.org/10.1007/s40319-023-01399-4>
- Spindler, G. (2019). Copyright law and artificial intelligence: G. Spindler. *IIC - International Review of Intellectual Property and Competition Law*, 50(9), 1049-1051. <https://doi.org/10.1007/s40319-019-00879-w>
- Szkalej, K. y Senftleben, M. (2024). Generative AI and creative commons licences: The application of share alike obligations to trained models, curated datasets and AI output. *J. Intell. Prop. Info. Tech. & Elec. Com. L.*, 15, 313. <https://www.jipitec.eu/jipitec/article/view/415>
- Thambaiya, N., Kariyawasam, K. y Talagala, C. (2025). Copyright law in the age of AI: Analysing the AI-generated works and copyright challenges in Australia. *International Review of Law, Computers & Technology*, 39(3), 448-473. <https://doi.org/10.1080/13600869.2025.2486893>
- Ueno, T. (2021). The flexible copyright exception for “non-enjoyment” purposes – Recent amendments in Japan and its implication. *GRUR International*, 70(2), 145-152. <https://doi.org/10.1093/grurint/ikaa184>
- Vincent, J. (15 de marzo de 2023). OpenAI co-founder on company’s past approach to openly sharing research: “We were wrong”. *The Verge*. <https://tinyurl.com/ecp69dv9>
- World Intellectual Property Organization. (20 de diciembre de 1996). *Agreed statements concerning the WIPO Copyright Treaty*. TRT/WCT/002. <https://tinyurl.com/bde54zf6>
- Yan, X., Xiao, Y. y Jin, Y. (2024). Generative large language models explained [AI-eXplained]. *IEEE Computational Intelligence Magazine*, 19(4), 45-46. <https://doi.org/10.1109/MCI.2024.3431454>
- Zhou, Z., Xiang, J., Chen, C. y Su, S. (2024). Quantifying and analyzing entity-level memorization in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17), 19741-19749. <https://doi.org/10.1609/aaai.v38i17.29948>
- Ziaja, G. M. (2024). The text and data mining opt-out in Article 4(3) CDSMD: Adequate veto right for rightholders or a suffocating blanket for European artificial intelligence innovations? *Journal of Intellectual Property Law & Practice*, 10, 453-464. <https://doi.org/10.1093/jiplp/jpae025>

Legislación citada

- Acuerdo sobre los Aspectos de los Derechos de Propiedad Intelectual relacionados con el Comercio [ADPIC], 15 de abril de 1994, Marrakesh Agreement Establishing the World Trade Organization, Annex 1C, 1869 U.N.T.S. 299.
- California AI Transparency Act, S.B. 942, 2023-2024 Leg. (Cal. 2024).
- Convenio de Berna para la Protección de las Obras Literarias y Artísticas, 9 de septiembre de 1886 (revisado en París el 24 de julio de 1971).
- Directiva 2019/790 del Parlamento Europeo y del Consejo, de 17 de abril de 2019, sobre los derechos de autor y derechos afines en el mercado único digital y por la que se modifican las Directivas 96/9/CE y 2001/29/CE. *Diario Oficial de la Unión Europea*, L 130/92.
- Directiva 2001/29/CE del Parlamento Europeo y del Consejo, de 22 de mayo de 2001, relativa a la armonización de determinados aspectos de los derechos de autor y derechos afines a la propiedad intelectual en la sociedad de la información. *Diario Oficial de las Comunidades Europeas*, L 167/10.
- Reglamento 2024/1689/UE del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial. *Diario Oficial de la Unión Europea*, L 2024/1689.

Jurisprudencia citada

- Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015).
- Authors Guild v. HathiTrust*, 755 F.3d 87 (2d Cir. 2014).
- Campbell v. Acuff-Rose Music, Inc.*, 510 U.S. 569 (1994).
- The New York Times Company v. OpenAI LP*, No. 23-cv-11222 (S.D.N.Y. filed Dec. 27, 2023).

* * * *

Roles de autoría y conflicto de intereses

El autor manifiesta que cumplió todos los roles de autoría del presente artículo y declara no poseer conflicto de interés alguno.