S

OMPI
ORGANIZACIÓN MUNDIAL
DE LA PROPIEDAD
INTELECTUAL

SCCR/49/6
ORIGINAL: INGLÉS
FECHA: 23 DE OCTUBRE DE 2026

Comité Permanente de Derecho de Autor y Derechos Conexos

Cuadragésima novena sesión
Ginebra, 30 de noviembre a 4 de diciembre de 2026

ESTUDIO EXPLORATORIO SOBRE EL ENTRENAMIENTO DE LA INTELIGENCIA ARTIFICIAL Y EL DERECHO DE AUTOR

preparado por Daniel J. Gervais, Sung Jai Choi, Melissa Omino, Rodrigo Kopke Salinas, Wang Qian

Estudio exploratorio sobre el entrenamiento de la IA y el derecho de autor

DESCARGO DE RESPONSABILIDAD:
Este estudio fue encargado por la Secretaría de la Organización Mundial de la Propiedad Intelectual (OMPI). Las opiniones expresadas en el presente documento son responsabilidad exclusiva de sus autores y no se pretende que reflejen las opiniones de los Estados miembros ni de la Secretaría de la OMPI.

AGRADECIMIENTOS

DANIEL J. GERVAIS es el coautor principal de este estudio. Es catedrático emérito de Derecho (Milton R. Underwood) en la Facultad de Derecho de Vanderbilt, profesor asociado superior en la Vrije Universiteit de Ámsterdam y miembro electo del American Law Institute. Sus investigaciones están disponibles en www.collectivemanagement.net, www.tripsagreement.com y www.thehumancause.org.

SUNG JAI CHOI es catedrático de Derecho en la Universidad de Sejong. Posee un máster en Derecho (LL.M.) por la Facultad de Derecho de Columbia. Su investigación se centra en la propiedad intelectual y el derecho de la competencia. Es coautor de la Ley de Derecho de Autor (2024) y recibió el Gran Premio de Arbitraje de Corea (2021) en la categoría de “Líderes de la próxima generación”.

MELISSA OMINO es directora del Centro de Derecho de la Propiedad Intelectual y las Tecnologías de la Información (CIPIT) de la Universidad de Strathmore en Nairobi. Es doctora en propiedad intelectual y derecho mercantil y está especializada en propiedad intelectual, inteligencia artificial, protección de datos y derechos digitales en el Sur Global.

RODRIGO KOPKE SALINAS es profesor de la Facultad de Derecho de la FGV de São Paulo (GVlaw) y cuenta con un máster en Derecho (LL.M.) por la Universidad de Columbia. Es miembro del Consejo Especial de Derecho de Autor del Colegio de Abogados del Brasil y está especializado en derecho de los medios de comunicación y del entretenimiento.

WANG QIAN es catedrático de Derecho y vicepresidente de la Comisión Académica de la East China University of Political Science and Law. Además, es vicepresidente de la Asociación China de Derecho de Autor y formó parte del Comité de Expertos encargado de la tercera revisión de la Ley de Derecho de Autor de China.

NOTA: Se han utilizado modelos de lenguaje para revisar la sintaxis, la gramática y el contenido del estudio, así como para aportar sugerencias editoriales con el fin de armonizar este texto escrito en colaboración.

ÍNDICE

Introducción....4
Parte I. Uso de contenidos protegidos en el desarrollo de modelos de IA generativa....5
Sección A. Obtención de los datos....5
1. Patrones de consumo de datos por tipo de sistema....5
2. Raspado web....6
3. Generación aumentada por recuperación....7
Sección B. Uso de los datos en el entrenamiento....8
1. Entrenamiento del modelo....9
2. Optimización y aprendizaje por refuerzo....10
Sección C. Tras el entrenamiento....11
1. Otros usos del corpus de entrenamiento....11
2. Conservación de copias en el modelo....11
3. La relación entre los datos de entrada y los de salida....15
Parte II. Autorización, concesión de licencias y compensación....16
Sección A. Opciones actuales de los desarrolladores para obtener materia susceptible de protección por derecho de autor con fines de entrenamiento....17
1. Las cinco vías de adquisición....17
2. Acuerdos de colaboración....19
Sección B. Principales elementos de política....19
1. Derechos afectados por el entrenamiento de los modelos de IA generativa....19
a) Derecho de reproducción....19
b) El derecho de adaptación y las obras derivadas....20
c) Comunicación al público y puesta a disposición....21
d) Derechos morales....21
e) Información sobre la gestión de derechos y medidas contra la elusión....22
2. Concesión voluntaria de licencias (colectivas e individuales)....22
a) Concesión de licencias individuales y colectivas para el uso de la IA....22
b) Principales modelos de licencias colectivas (voluntarias, ampliadas y obligatorias)....27
3. Licencias comunitarias y de custodia para datos lingüísticos primarios....31
4. Excepciones legales relativas a la minería de textos y datos y cláusulas generales....32
5. Los sistemas de remuneración....36
a) Tipos de sistemas de remuneración....36
b) Modalidades y base de la remuneración....38
Parte III. Observancia....39
Sección A. Enseñanzas de la jurisprudencia actual....40
1. La cuestión del “modelo como copia”....40
2. La excepción relativa a la minería de textos y datos....42

3. Competencia respecto del entrenamiento extraterritorial 43
4. La sustitución 45
Sección B. Observancia frente a los desarrolladores de IA 49
1. Rasgos estructurales del panorama de la observancia 49
2. Asimetrías entre las partes 50
Sección C. Desafíos en la observancia de los derechos 50
1. Transparencia 51
2. Ataques a los datos, filtrado y proliferación de normas de exclusión voluntaria 52
3. La jurisprudencia en su conjunto 56
Conclusión 57

INTRODUCCIÓN

El presente estudio se ha elaborado a petición del Comité Permanente de Derecho de Autor y Derechos Conexos (SCCR). En su cuadragésima séptima sesión (celebrada en Ginebra del 1 al 5 de diciembre de 2025), el Comité pidió a la Secretaría que prepare un estudio sobre los enfoques normativos o reglamentarios de la relación entre el entrenamiento de la IA y el derecho de autor, y las normas conexas, así como las prácticas aplicables en materia de autorización, observancia y compensación por el uso. La solicitud figura en el resumen de la presidencia de dicha sesión (documento SCCR/47/SUMMARY[1], de 5 de diciembre de 2025, párrafo 21).

El estudio es de carácter descriptivo y comparativo. En él se analizan los métodos adoptados en una serie de jurisdicciones y las prácticas que se están desarrollando, y se examinan las opciones que se han propuesto, sin recomendar ningún enfoque concreto a los Estados miembros ni pronunciarse sobre las cuestiones objeto de controversia que aún deben resolver.

El estudio trata sobre el entrenamiento de modelos de IA generativa, esto es, el uso de contenidos para la creación y la mejora de dichos modelos, así como sobre las cuestiones relativas a la autorización, la observancia y la compensación que plantea dicho uso cuando el contenido está amparado por el derecho de autor o derechos conexos. La cuestión más amplia de la relación entre la inteligencia artificial y el derecho de autor va mucho más allá del entrenamiento. En particular, aborda la situación jurídica de los resultados generados por un modelo desplegado, tanto el hecho de si un resultado concreto puede infringir el derecho de autor de las obras que se utilizaron para entrenarlo como el hecho de si un resultado puede estar protegido por derecho de autor y en qué condiciones. Las preguntas relativas a los resultados plantean otras cuestiones que el Comité no pidió que se abordaran en este estudio. Solo se mencionan cuando es necesario para ofrecer una visión completa de la cuestión del entrenamiento o para dar todo el contexto.

El estudio se presenta en tres partes. En la parte I se describen, desde un punto de vista técnico y descriptivo, las operaciones mediante las cuales se utiliza contenido protegido para desarrollar un modelo de IA generativa y se identifican los aspectos en los que dichas operaciones pueden afectar a los derechos exclusivos reconocidos por la legislación nacional y los tratados internacionales en materia de derecho de autor. En la parte II se analizan el alcance y la extensión de los derechos en cuestión, así como los mecanismos mediante los cuales estos se ejercen: concesión de licencias voluntarias, individuales o colectivas; y las excepciones y limitaciones aplicables. La parte III aborda la observancia y analiza las enseñanzas que pueden extraerse del corpus de jurisprudencia que ha comenzado a elaborarse en varias jurisdicciones. Asimismo, aborda el carácter transfronterizo del entrenamiento y las cuestiones de observancia conexas, entre otras la transparencia y los protocolos para manifestar la exclusión voluntaria, cuando proceda. Las jurisdicciones citadas en este estudio se mencionan a título ilustrativo y no constituyen una lista exhaustiva. Entre otras se encuentran las del Brasil, el Canadá, China, los Estados Unidos, Kenya, la Unión Europea y muchos de sus Estados miembros, el Reino Unido y la República de Corea.

[1] Véase https://www.wipo.int/meetings/es/doc_details.jsp?doc_id=653196

PARTE I. USO DE CONTENIDOS PROTEGIDOS EN EL DESARROLLO DE MODELOS DE IA GENERATIVA

Un modelo de inteligencia artificial (IA) generativa se desarrolla en etapas sucesivas y, por lo general, en cada una de ellas se utiliza contenido de terceros. Parte del contenido puede estar protegido por derecho de autor y por derechos conexos; otra parte puede pertenecer al dominio público, ya sea porque ha expirado el plazo de protección o porque nunca estuvo protegido. Las operaciones técnicas que se aplican a dicho contenido suelen ser las mismas, independientemente de su situación jurídica. Los mismos procesos, entre otros el rastreo, la copia, la *tokenización* y el análisis estadístico (los cuales se explican a continuación), pueden aplicarse del mismo modo tanto a una novela reciente como a un texto del siglo XVIII que forma parte del dominio público. En esta parte se describen dichas operaciones y se señalan, a nivel descriptivo, los puntos en los que pueden afectar a derechos reconocidos por las legislaciones nacionales y por los tratados internacionales sobre derecho de autor y derechos conexos.

A continuación se presentan las tres fases del entrenamiento. En la sección A se aborda la adquisición de datos, es decir, cómo el desarrollador de IA crea el corpus, a menudo compuesto en gran parte por material protegido, sobre el que se entrenará un modelo. La sección B aborda el uso real de esos datos en el entrenamiento, incluyendo la sesión principal de entrenamiento y los distintos tipos de ajuste fino posterior. En la sección C se analiza la situación tras el entrenamiento del modelo, incluido el tratamiento del corpus de entrenamiento, así como la cuestión de si el propio modelo entrenado contiene copias de las obras[2] con las que se ha entrenado.

SECCIÓN A. OBTENCIÓN DE LOS DATOS

El desarrollo de un gran modelo de lenguaje comienza con la creación de un corpus, normalmente a gran escala y de procedencia variada. Los desarrolladores obtienen los datos de varias formas. Algunos datos pueden estar sujetos a una licencia; en tal caso, el contenido podrá facilitarse junto con la licencia para su uso. En ese supuesto, un desarrollador puede adquirir el acceso a un conjunto de datos seleccionado y recopilado por un tercero, o negociar un acuerdo de licencia directo con una editorial, un medio de comunicación o un organismo de gestión colectiva (OGC). Si no se obtiene ninguna licencia para el uso de las obras protegidas, dicho uso quizá se base en la suposición de que se aplica una excepción o limitación al derecho exclusivo correspondiente. Una vía de entrenamiento muy importante, tanto por su volumen como por las cuestiones de derecho de autor que plantea, es la recuperación automatizada de material de la web abierta (Internet). El entrenamiento también puede recurrir a datos sintéticos, es decir, datos generados por un modelo de IA ya existente en lugar de datos recopilados de creadores humanos, especialmente en las últimas fases del desarrollo.

1. Patrones de consumo de datos por tipo de sistema

La demanda de datos de un sistema no suele limitarse a la sesión de entrenamiento inicial: los datos se utilizan para el reentrenamiento y ajuste fino, para el análisis de texto y datos que se lleva a cabo en una biblioteca de búsqueda actualizada, para la generación aumentada por recuperación (RAG)[3] y para la generación de vistas previas y fragmentos. Estos usos varían en cuanto a la cantidad de material que requieren y a la facilidad con la que se pueden rastrear sus fuentes. El tipo de sistema suele influir en los patrones de consumo: los sistemas de generación de contenidos multimedia y los *chatbots* basados en grandes modelos de lenguaje consumen datos principalmente durante la ingestión de datos de entrenamiento; los sistemas

[2] Para simplificar el texto, debe entenderse que el término “obras” incluye los objetos de derechos conexos cuando proceda.

[3] La generación aumentada por recuperación, conocida comúnmente por sus siglas RAG, se analiza con más detalle en la sección 3 de esta parte.

personalizables creados por las empresas a partir de un modelo de lenguaje (también llamado modelo fundacional) añaden un corpus específico del ámbito y, a menudo, una biblioteca de búsqueda a la que se recurre en el momento de su uso; y los motores de búsqueda conversacionales requieren acceso a fuentes en línea en el momento de la generación. La posición que ocupa un sistema en este espectro influye en el conjunto de derechos exclusivos que conlleva el uso de los datos y en la estructura correspondiente de cualquier licencia. Esta cuestión se desarrolla en la parte II, sección B.2.

2. Raspado web

El "raspado web" (*web scraping*) es un método habitual que utilizan los desarrolladores de modelos fundacionales para crear sus corpus de entrenamiento. Ahora bien, esta terminología no se ha adoptado de forma generalizada. Algunas fuentes distinguen entre el "rastreo web" (*web crawling*) (el recorrido automatizado por sitios web que realiza un programa siguiendo los enlaces de una página a otra) y el "raspado web" (la obtención selectiva de material identificado de una página que ya se ha recuperado). En la práctica, ambas actividades pueden combinarse: un desarrollador despliega un rastreador que sigue los enlaces de la web y, en cada página, identifica y extrae el contenido textual, visual o de otro tipo que se desea conservar. El resultado suele ser una copia de ese contenido, que normalmente se almacena de forma estructurada en la propia infraestructura del desarrollador.

Hay cuatro aspectos que son relevantes para el presente análisis. En primer lugar, la técnica consiste en realizar copias. Las copias pueden almacenarse en un servidor o en una memoria caché, o bien procesarse en la memoria antes de someterse a una transformación posterior. Estas operaciones pueden afectar al derecho de reproducción reconocido en el artículo 9.1) del Convenio de Berna, en el artículo 1.4) del Tratado de la OMPI sobre Derecho de Autor (WCT) y en las disposiciones correspondientes de las legislaciones nacionales. En la declaración concertada sobre el artículo 1.4) del WCT se confirma que el almacenamiento de una obra protegida en formato digital constituye una reproducción y que el derecho de reproducción y sus excepciones se aplican plenamente en el entorno digital.[4] La cuestión de si el carácter temporal de determinadas copias, o la aplicación de una excepción como la relativa a reproducciones incidentales y transitorias, modifica el análisis se aborda más adelante en este estudio. El artículo 7 del Tratado de la OMPI sobre Interpretación o Ejecución y Fonogramas (WPPT) y el mismo artículo del Tratado de Beijing sobre Interpretaciones y Ejecuciones Audiovisuales también recogen un derecho relativo a la reproducción directa o indirecta "por cualquier procedimiento o bajo cualquier forma".

En segundo lugar, la recuperación puede ser no selectiva. Un rastreador no siempre distingue entre el contenido que pertenece al dominio público, el contenido que se ofrece bajo una licencia abierta, el contenido que se pone a disposición sin ninguna indicación de licencia y el contenido que está protegido y se pone a disposición con arreglo a unas condiciones de uso o medidas tecnológicas de protección específicas. Por ejemplo, algunos rastreadores, aunque no todos, respetan las instrucciones establecidas en los archivos robots.txt.[5] A veces, estas distinciones se establecen *a posteriori*, mediante el filtrado o la eliminación de duplicados del corpus, pero el acto de copiado inicial no siempre es selectivo.

[4] Tratado de la OMPI sobre Derecho de Autor (20 de diciembre de 1996), Declaración concertada respecto del artículo 1.4: "El derecho de reproducción, tal como se establece en el Artículo 9 del Convenio de Berna, y las excepciones permitidas en virtud del mismo, son totalmente aplicables en el entorno digital, en particular a la utilización de obras en forma digital. Queda entendido que el almacenamiento en forma digital en un soporte electrónico de una obra protegida, constituye una reproducción en el sentido del Artículo 9 del Convenio de Berna."

[5] El protocolo de exclusión de robots, normalizado como RFC 9309 (2022) del Grupo de Trabajo de Ingeniería de Internet (IETF), es una convención mediante la cual el administrador de un sitio web coloca un archivo de texto sin formato (robots.txt) en la raíz del sitio para indicar qué rastreadores automáticos pueden acceder a qué partes del mismo. El cumplimiento es voluntario; el protocolo no cuenta con ningún mecanismo de observancia.

En tercer lugar, la visibilidad de la actividad para los titulares de derechos es limitada. Es posible que el titular de derechos no sepa que su obra ha sido "raspada", ni por qué rastreador, cuándo ni para qué uso posterior. Las obligaciones de divulgación que están empezando a aparecer en algunas legislaciones nacionales, en particular en la Unión Europea en virtud del Reglamento de Inteligencia Artificial de 2024 (Reglamento de IA), exigen a los proveedores de modelos de IA de uso general que publiquen un resumen "suficientemente detallado" del contenido utilizado para el entrenamiento.[6] Aún es pronto para evaluar el efecto práctico de esas obligaciones. La cuestión de cómo se puede conciliar la divulgación de información con el respeto a los intereses comerciales legítimos de los desarrolladores es, en sí misma, objeto de un debate en curso.[7]

En cuarto lugar, la realización y el uso de copias de obras protegidas con fines de entrenamiento pueden implicar la eliminación o la alteración de la información sobre la gestión de derechos, así como la elusión de medidas tecnológicas efectivas que utilizan los titulares para hacer valer sus derechos en virtud del WCT, el WPPT y el Tratado de Beijing.

3. Generación aumentada por recuperación

Una técnica distinta, aunque relacionada, es la generación aumentada por recuperación (RAG).[8] Se trata de un método mediante el cual un sistema de IA responde a una consulta consultando primero una colección independiente y previamente recopilada de documentos, en lugar de basarse únicamente en el material con el que se entrenó. Una comparación útil es la diferencia entre una persona que responde a una pregunta de memoria y esa misma persona que se detiene para consultar algo en una estantería de referencia situada junto al escritorio. Cuando llega una consulta, el sistema busca en esa colección, extrae los fragmentos más relevantes para la pregunta, los coloca junto a la consulta y los utiliza para elaborar su respuesta. La colección se elabora con antelación mediante la creación y el almacenamiento de copias de los documentos, y los pasajes pertinentes suelen copiarse de nuevo en el sistema en el momento en que se genera la respuesta, aspectos que son relevantes para el análisis relativo al derecho de autor que se expone a continuación.

En un sistema RAG, se mantiene un corpus externo, que suele indexarse en una base de datos después de que cada documento o fragmento de documento se ha convertido en una representación vectorial.[9] Cuando el sistema recibe la consulta de un usuario, esta se transforma en una representación vectorial y se utiliza para recuperar los fragmentos más relevantes del corpus. A continuación, el sistema elabora el *prompt* (la instrucción) que se envía al modelo de lenguaje, combinando la consulta del usuario con los fragmentos recuperados,

[6] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial), artículo 53, apartado 1, letra d) (obligación de los proveedores de modelos de IA de uso general de elaborar y poner a disposición del público un resumen suficientemente detallado del contenido utilizado para el entrenamiento del modelo de IA de uso general, con arreglo al modelo facilitado por la Oficina de IA).

[7] Sin embargo, en ocasiones, el origen del material obtenido mediante raspado web se refleja en el comportamiento del modelo resultante: las primeras versiones de un destacado modelo de generación de imágenes reproducían una forma reconocible de la marca de agua de una agencia de imágenes de archivo, lo que indicaba que se había incorporado sin autorización un gran volumen de imágenes de dicha agencia. El fenómeno fue ampliamente difundido y se abordó en el litigio analizado en la sección A.1 de la parte III (*Getty Images (US) Inc. v. Stability AI Ltd.*). Esto pone de manifiesto que, si bien el raspado de una obra concreta suele pasar desapercibido para el titular de los derechos, el propio modelo entrenado puede revelar, a través de sus resultados, que el material protegido se utilizó durante el entrenamiento.

[8] Véase la nota 2 *supra*.

[9] Las representaciones vectoriales de palabras asignan a cada *token* un vector denso de alta dimensión compuesto por valores reales. Las representaciones vectoriales se describen como "contextuales" cuando la representación asignada a un *token* determinado depende de los *tokens* que lo rodean. Véase "The Heart of the Matter", nota 10 *supra*, en la parte II.C.

que el modelo utiliza para generar una respuesta basada en su entrenamiento y en el material recuperado.

A los efectos que nos ocupan, el RAG en ocasiones se describe como un sistema que repercute en las cuestiones de derecho de autor que surgen del entrenamiento de los modelos de lenguaje. El razonamiento es que el modelo en sí mismo no aprende del corpus, ya que este se mantiene por separado y se consulta en la fase de inferencia. La creación del corpus indexado requiere un procesamiento previo y, en una arquitectura RAG convencional, el almacenamiento de los documentos originales o de los fragmentos que los componen.[10] La conversión de cada documento en una representación vectorial da lugar a una representación numérica derivada del documento; sin embargo, a diferencia de la copia original almacenada, la representación vectorial no suele ser reversible al texto original. Durante la fase de inferencia (ejecución), los fragmentos pertinentes se copian en la ventana de contexto del modelo y pueden reproducirse en el resultado generado, en su totalidad o en una parte sustancial. Estos pasos afectan al derecho de reproducción y podrían afectar al derecho de adaptación (artículo 12 del Convenio de Berna) dependiendo de la legislación aplicable y de las características de la representación. La posible aplicación de dichos derechos se analiza más adelante.

SECCIÓN B. USO DE LOS DATOS EN EL ENTRENAMIENTO

El entrenamiento de un modelo de IA generativa es una secuencia de operaciones que normalmente se describen en cuatro etapas: preentrenamiento, ajuste fino supervisado, modelado de recompensas y aprendizaje por refuerzo.[11] La primera de ellas suele ser la más exigente, tanto en términos de costo computacional como de volumen de datos que consume; las tres etapas posteriores, en su conjunto, suelen describirse como “ajuste fino” (*fine-tuning*) o, en ocasiones, como “posentrenamiento”. Una actividad distinta pero relacionada, consistente en la evaluación y optimización del modelo, acompaña al entrenamiento y continúa una vez desplegado el modelo.

A efectos del derecho de autor, la observación principal es que cada una de estas etapas suele implicar el almacenamiento y el tratamiento de la totalidad o de parte del corpus subyacente. El proceso de entrenamiento no consiste en consultar los datos brevemente para luego descartarlos; los datos se almacenan, se cargan, se transforman y se utilizan para actualizar el modelo. Cuando los datos incluyen obras protegidas, dichas operaciones implican la

---

[10] En términos más técnicos, conviene distinguir entre las representaciones vectoriales de fragmentos de texto utilizadas para la búsqueda vectorial y el texto original a partir del cual se generan. Una representación vectorial es una representación numérica densa generada por un codificador; no es una copia exacta del texto y, por lo general, no es reversible simplemente descodificando el vector. No obstante, las técnicas especializadas de inversión de “representaciones vectoriales” pueden reconstruir algunos o, en condiciones experimentales específicas, la totalidad de fragmentos relativamente breves de la fuente a partir de sus representaciones vectoriales. En un sistema RAG convencional, los documentos o fragmentos de origen también se conservan en formato de texto, ya sea en el propio almacén de vectores o en un almacén de documentos independiente vinculado a los vectores. Las representaciones vectoriales se utilizan para identificar fragmentos relevantes; a continuación, se recupera el texto asociado a ellas, y no el correspondiente vector de representación, y se proporciona al modelo de lenguaje como parte de su contexto de entrada. Véase John X. Morris *et al.*, “Text Embeddings Reveal (Almost) As Much As Text”, Actas de la Conferencia de 2023 sobre Métodos Empíricos en el Procesamiento del Lenguaje Natural (EMNLP). Para una descripción del proceso de recuperación desde la perspectiva del sector, véase: https://infohub.delltechnologies.com/en-us/p/demystifying-on-device-intelligent-search-using-rag-architecture/

[11] La descripción en cuatro etapas que se utiliza en el presente documento procede de la bibliografía técnica y se analiza con mayor detalle en el artículo “The Heart of the Matter: Copyright, AI Training, and LLMs” 71 *J Cop Soc’y of the USA* 482 (2024) (en adelante, “The Heart of the Matter”), disponible en https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4963711. En la página 10 de la *Guía sobre la aplicabilidad de la doctrina del uso legítimo al entrenamiento de modelos de IA generativa* (febrero de 2026), publicada por el Ministerio de Cultura, Deporte y Turismo y la Comisión de Derecho de Autor de Corea (en adelante, la “Guía de Corea sobre el uso legítimo”), se presenta un proceso similar de cuatro etapas (recopilación de datos, preprocesamiento de datos, entrenamiento del modelo y evaluación y optimización del modelo).

realización de copias o transformaciones relevantes para el análisis de la reproducción que se desarrolla en la parte II.

1. Entrenamiento del modelo

La principal fase de entrenamiento, denominada generalmente "preentrenamiento", es la etapa en la que el modelo adquiere la mayor parte de su conocimiento sobre el lenguaje (en el caso de un modelo de lenguaje) o sobre imágenes, audio y video (en el caso de un modelo multimodal). El preentrenamiento se lleva a cabo en varias etapas técnicas, cada una de las cuales se realiza sobre una copia del corpus almacenada. En primer lugar, el contenido textual se somete a un proceso de *tokenización*, que consiste en convertir palabras y otras secuencias en una secuencia de unidades básicas de información numéricas (*tokens*) extraídas de un vocabulario fijo.[12] Los *tokenizadores* modernos suelen funcionar a nivel de subpalabra y no de palabras completas.[13] Este paso es relevante para el presente análisis porque, en el caso de los *tokenizadores* a nivel de *byte* que utilizan los principales modelos actuales, es un proceso sin pérdida de información y reversible: el texto codificado por un *tokenizador* de este tipo y posteriormente descodificado da como resultado el texto original, por lo que el corpus *tokenizado* no es un resumen ni una abstracción de las obras, sino una recodificación de las mismas en otro formato.

A continuación, los *tokens* se convierten en representaciones vectoriales.[14] En el enfoque que ha dominado este campo técnico desde 2017, el "Transformador" (la "T" de las siglas GPT, que significan *Generative Pre-trained Transformer* o "transformador generativo preentrenado"), se procesan esas representaciones mediante un mecanismo conocido como "atención".[15] La caracterización de dicha representación vectorial y del propio modelo de entrenamiento, que implican un mayor grado de transformación, se aborda en la sección C.2 más adelante.

Una vez establecidos los parámetros iniciales del modelo, se realiza el entrenamiento sometiéndolo repetidamente a una secuencia de *tokens* extraídos del corpus. El modelo predice cada *token* sucesivamente basándose en los *tokens* que le preceden y sus parámetros se van ajustando gradualmente para que sus predicciones se acerquen cada vez más al texto observado; de este modo, el corpus no se consulta una única vez, sino que se utiliza, *token* a *token*, a lo largo de todo el proceso de entrenamiento. Tras procesar un volumen enorme de *tokens* extraídos del corpus, el modelo aprende una inmensa distribución de probabilidades sobre un vasto espacio de posibles secuencias de *tokens*. Esa distribución, codificada en los parámetros del modelo, es lo que permite que este genere texto coherente en respuesta a una instrucción o *prompt*.

La relación entre los datos de entrenamiento y el modelo entrenado es objeto de un intenso debate. Una de las posturas es que, dado que el modelo entrenado solo contiene parámetros y no los datos de entrenamiento originales, no queda ninguna copia de las obras protegidas una

[12] Sobre la *tokenización*, véase "The Heart of the Matter", nota 10 *supra*, en la parte II.B. La Guía de Corea sobre el uso legítimo se refiere a la *tokenización* como uno de los procedimientos técnicos de preprocesamiento de datos, junto con la limpieza y la normalización. Véase la Guía de Corea sobre el uso legítimo, nota 10 *supra*, pág. 10.

[13] El vocabulario se construye partiendo de caracteres individuales y fusionando progresivamente los pares adyacentes que aparecen con mayor frecuencia para formar unidades más grandes, de modo que las palabras comunes quedan representadas como un único *token*, mientras que las menos frecuentes se dividen en varios. OpenAI utiliza la codificación por pares de *bytes* en sus modelos GPT (transformadores generativos preentrenados), al igual que otros importantes desarrolladores. Véase "The Heart of the Matter", nota 10 *supra*, en la parte II.B.

[14] Más concretamente, se proyectan en un espacio vectorial de alta dimensión, en el que cada uno de ellos se convierte en una representación vectorial.

[15] El mecanismo de atención permite al modelo, al calcular la representación de un *token* determinado, valorar la relevancia de todos los demás *tokens* de la secuencia, lo que le permite captar dependencias a lo largo de pasajes extensos de texto. Ashish Vaswani *et al.*, "Attention Is All You Need", 31.ª Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS) (2017). La arquitectura del "Transformador" es la base de los principales grandes modelos de lenguaje que existen en la actualidad.

vez finalizado el entrenamiento.[16] Los parámetros describen una distribución de probabilidades en un espacio vectorial enormemente extenso. Si bien esta distribución permite al modelo generalizar y generar secuencias totalmente nuevas, puede asignar probabilidades desproporcionadamente altas y concentradas a secuencias específicas y exactas que se han encontrado repetidamente durante el entrenamiento. Dado que la distribución atribuye una probabilidad especialmente elevada a esas secuencias finitas presentes en los datos de entrenamiento, el modelo tenderá, en muchos casos, a reproducirlas de forma literal o casi literal. Este fenómeno está ampliamente documentado en la bibliografía técnica y se analiza con más detalle en la sección C que figura más adelante. Como se explica más adelante, existe una diferencia entre este fenómeno (al que a menudo se denomina "memorización") y la cuestión de si existe una "reproducción" dentro del modelo de todas las obras utilizadas para el entrenamiento.

Respecto a los modelos de generación de imágenes, audio o video, los modelos de "difusión" son la clase predominante; se entrenan añadiendo progresivamente ruido a los ejemplos de entrenamiento y aprendiendo posteriormente a revertir ese proceso para generar nuevos ejemplos a partir del ruido.[17] Esto difiere de los modelos de lenguaje (texto), pero en ambos casos el proceso de entrenamiento genera una representación sofisticada de los datos de entrada.

2. Optimización y aprendizaje por refuerzo

Las etapas posteriores al preentrenamiento son las encargadas de adaptar el modelo preentrenado a tareas concretas, casos de uso específicos y determinadas pautas de comportamiento. La terminología varía según los desarrolladores, pero la secuencia incluye el ajuste fino supervisado, el modelado de recompensas y el aprendizaje por refuerzo.[18]

El ajuste fino supervisado consiste en continuar el entrenamiento con un conjunto de datos más reducido y cuidadosamente seleccionado, compuesto por pares de entrada y salida del tipo que se espera que produzca el modelo. Ese conjunto de datos es mucho más pequeño que el corpus de preentrenamiento, pero se caracteriza por su calidad y, a menudo, ha sido elaborado o revisado por humanos. Cada ejemplo de ajuste fino, al igual que cada ejemplo de preentrenamiento, se carga y se utiliza para actualizar los parámetros del modelo.

El modelado de recompensas consiste en la construcción de un modelo "de recompensas" independiente, que asigna una puntuación a los posibles resultados del modelo principal en función de su grado de adecuación a los resultados esperados. El modelo de recompensas se entrena a partir de comparaciones generadas por personas entre los resultados candidatos.

---

[16] Véase "The Heart of the Matter", nota 10 *supra*, en la parte II.C.

[17] Sobre los orígenes de los modelos de difusión, véase Jascha Sohl-Dickstein et al., "Deep Unsupervised Learning using Nonequilibrium Thermodynamics", Actas de la 32.ª Conferencia Internacional sobre Aprendizaje Automático (ICML) (2015); Jonathan Ho, Ajay Jain y Pieter Abbeel, "Denoising Diffusion Probabilistic Models", 34.ª Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS) (2020); y, respecto a la arquitectura de difusión latente utilizada en Stable Diffusion y objeto de controversia en el litigio con Getty analizado en la parte III, Robin Rombach et al., "High-Resolution Image Synthesis with Latent Diffusion Models", Conferencia IEEE/CVF sobre Visión por Ordenador y Reconocimiento de Patrones (CVPR) (2022). Para una introducción accesible, véase también Yang Song y Stefano Ermon, "Generative Modeling by Estimating Gradients of the Data Distribution", 33.ª Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS) (2019), disponible en http://papers.neurips.cc/paper/9361-generative-modeling-by-estimating-gradients-of-the-data-distribution.pdf. La cuestión general examinada en el presente debate no depende de los detalles matemáticos de ninguna clase concreta de modelo.

[18] Véase "The Heart of the Matter", nota 10 *supra*, en la parte II.D. Términos como "preentrenamiento", "ajuste fino supervisado", "modelado de recompensas" y "aprendizaje por refuerzo" se utilizan de forma generalizada entre los desarrolladores de IA.

El aprendizaje por refuerzo a partir de comentarios humanos (RLHF, por sus siglas en inglés), una técnica muy utilizada para ajustar los modelos de IA generativa a las preferencias humanas, utiliza el modelo de recompensas para orientar las modificaciones posteriores del modelo principal.

Cada una de estas etapas requiere el almacenamiento y el procesamiento del material de entrenamiento. Los volúmenes son menores que los necesarios para el preentrenamiento, pero el material de entrenamiento puede incluir a su vez obras protegidas, sobre todo cuando las comparaciones proporcionadas por los anotadores humanos se basan en dichas obras o las incluyen. El hecho de centrarse más en la calidad del material que en su cantidad no altera el análisis desde la perspectiva del derecho de autor.

La evaluación y optimización del modelo, que se llevan a cabo en paralelo a las fases de entrenamiento y continúan tras el despliegue, implican el uso de conjuntos de datos de evaluación para valorar el rendimiento del modelo e identificar los aspectos en los que es necesario seguir entrenándolo. Los conjuntos de datos de evaluación también pueden incluir obras protegidas, sobre todo cuando los conjuntos de referencia se han elaborado a partir de corpus de prueba disponibles públicamente.

SECCIÓN C. TRAS EL ENTRENAMIENTO

Una vez finalizado el entrenamiento, surgen tres cuestiones distintas. La primera es si el corpus se conserva, quién lo conserva y con qué fines posteriores. La segunda es si el modelo contiene copias de las obras con las que se entrenó y, en caso afirmativo, en qué sentido. La tercera es si dicho resultado reproduce las obras que se utilizaron en el entrenamiento, las adapta o se relaciona con ellas de algún otro modo.

1. Otros usos del corpus de entrenamiento

El corpus de entrenamiento sigue siendo relevante una vez finalizada la fase de entrenamiento. El desarrollador puede conservarlo con varios fines: para evaluar futuras versiones del modelo, para realizar ajustes finos adicionales sobre subconjuntos específicos, para facilitar investigaciones en curso o para ponerlo a disposición de terceros bajo licencia o por otros medios. Se puede combinar con nuevos materiales recopilados para formar un corpus más amplio destinado a la próxima generación de modelos. Puede venderse o compartirse (por ejemplo, con organizaciones afiliadas). En algunos casos, se publican en el dominio público, bajo licencias permisivas, variantes del corpus, incluidos subconjuntos filtrados o sin duplicados.

Cada uno de esos usos posteriores puede implicar un acto de reproducción, distribución y puesta a disposición del público en el sentido de la normativa nacional e internacional pertinente. En otras palabras, un corpus que ha cumplido su finalidad de entrenamiento puede seguir utilizándose después durante años, y las normas que rigen su uso posterior coinciden con las que regían su compilación inicial.

2. Conservación de copias en el modelo

El hecho de que el propio modelo contenga “copias” de las obras con las que se ha entrenado es una de las cuestiones más complejas desde el punto de vista técnico que se abordan en esta parte y una de las que distinguen a la IA generativa de las tecnologías anteriores. Las posturas adoptadas por los desarrolladores de IA, por un lado, y por los titulares de derechos, por otro, han sido, hasta hace poco, diametralmente opuestas.[19] Los desarrolladores de IA

[19] La divergencia va más allá de las posiciones defendidas en los litigios y se extiende a las prioridades de diseño. Un estudio Delphi desarrollado en varias fases con especialistas en ética de la IA, especialistas jurídicos, profesionales de la IA aplicada a la música y artistas, llevado a cabo para validar un marco de diseño responsable

[Sigue la nota en la página siguiente]

suelen sostener que el modelo entrenado solo contiene parámetros, no las obras en sí mismas, y que no se conservan copias a efectos del derecho de autor. Los titulares de los derechos han argumentado que el modelo es, en realidad, una copia de todo el corpus. Los datos empíricos recopilados en los últimos dos años sugieren que ambas posturas tienden a simplificar en exceso la relación entre la memorización y el modelo. Además, independientemente de cómo se resuelva esta controversia, ello no afecta a los actos anteriores: la compilación del corpus y su uso para el entrenamiento implican actos de reproducción cuya calificación jurídica es independiente de que el propio modelo terminado contenga copias.

A este respecto, existen dos vías principales de análisis técnico. La memorización es una propiedad del modelo entrenado, que puede recuperar secuencias de entrenamiento específicas en los parámetros del modelo con la fidelidad suficiente. Dado que la estructura interna de un modelo no puede inspeccionarse fácilmente de forma directa, la memorización se estudia mediante la extracción, es decir, la capacidad demostrada por el modelo, tras recibir los *prompts* adecuados, de reproducir secuencias de sus datos de entrenamiento literal o casi literalmente. La recuperación se utiliza entonces como prueba de la memorización. El hecho de que un determinado procedimiento de extracción no tenga éxito no demuestra que una obra no haya sido memorizada. En un estudio reciente, realizado por Cooper *et al.*, se aplicó una técnica de extracción probabilística a 200 libros en 14 modelos de lenguaje de pesos abiertos y se concluyó que el grado de memorización variaba considerablemente tanto en función del modelo como del libro.[20] La mayoría de los modelos más grandes no memorizaban la mayor parte de los libros, ni en su totalidad ni en parte. Sin embargo, cuando hubo memorización, esta podía ser notable: el estudio señaló que un modelo de amplia difusión había memorizado casi por completo la primera novela de *Harry Potter* y *1984*, de George Orwell; en el caso de la primera, el estudio demostró que era posible generar el libro entero de forma casi literal a partir de un *prompt* inicial compuesto únicamente por los primeros tokens del primer capítulo.

No obstante, esas demostraciones empíricas de la memorización no están exentas de controversias. Los desarrolladores de IA han respondido argumentando que las técnicas de formulación de *prompts* necesarias para inducir reproducciones literales son artificiosas y están orientadas a forzar el sistema, no son representativas del uso que se hace de estos sistemas en la práctica y pueden constituir en sí mismas una manipulación o un uso indebido del producto que queda fuera de su uso habitual (y, en algunos casos, de las condiciones de servicio). En el litigio iniciado por *The New York Times* contra OpenAI, la parte demandada pidió que se desestimaran las alegaciones de infracción directa del derecho de autor

---

para sistemas generativos de música, puso de manifiesto una convergencia "proteccionista" entre los artistas y los especialistas jurídicos en cuestiones de soberanía de los datos, en contraste con el enfoque de los profesionales, centrado en la viabilidad, y concluyó que este tipo de marcos funcionan mejor como un mecanismo de mediación entre las prioridades divergentes de las distintas partes interesadas, incluido el equilibrio entre la transparencia de los modelos y los intereses en materia de propiedad intelectual, que como instrumentos destinados a imponer un consenso. J. de Berardinis *et al.*, “RAIM: Operationalising Trustworthy AI in Music Generation”, *IEEE Transactions on Technology and Society*, número especial *Ethical Innovation with/in Music Technology* (de próxima publicación, 2026), página web del marco de trabajo: https://casbahlab.github.io/raim-framework/ (consultado el 25 de agosto de 2026).

[20] Feder Cooper, Aaron Gokaslan, Ahmed Ahmed, Amy B. Cyphert, Christopher De Sa, Mark A. Lemley, Daniel E. Ho y Percy Liang, "Extracting Memorized Pieces of (Copyrighted) Books from Open-Weight Language Models", arXiv:2505.12546 (2025), versión del 30 de noviembre de 2025 (publicado también como *Stanford Public Law Working Paper / WVU College of Law Research Paper*, n.º 2025-005). Merece la pena citar íntegramente la manera en que los autores formularon la cuestión (pág. 1): "Los demandantes y los demandados en los litigios de derecho de autor relacionados con la IA generativa suelen hacer afirmaciones categóricas y contrapuestas sobre la medida en que los grandes modelos de lenguaje (LLM) han memorizado la expresión protegida de los demandantes. Basándonos tanto en el aprendizaje automático como en el derecho de autor, demostramos que estas posturas polarizadas simplifican excesivamente la relación entre la memorización y el derecho de autor". En cuanto a la proposición de que el contenido memorizado debe estar codificado en los parámetros del modelo porque "no puede estar en ningún otro sitio", véase también A. Feder Cooper y James Grimmelmann (citados en el mismo texto, pág. 25).

basándose, entre otras cosas, en que el demandante había utilizado lo que aquella calificó de *prompts* rebuscados y manipulados para provocar la regurgitación de fragmentos literales que este denunciaba.[21] En abril de 2025, el tribunal de distrito para el Distrito Sur de Nueva York se negó a desestimar las alegaciones por ese motivo y, desde entonces, la controversia se ha desplazado, en la fase de práctica de la prueba, de la cuestión de si la "regurgitación" puede provocarse de manera artificiosa a la de con qué frecuencia se produce espontáneamente en la práctica, lo que culminó en una resolución que ordenó la presentación de una muestra representativa de 20 millones de registros (*logs*) de interacciones reales de los usuarios con el modelo.[22] El litigio y la importancia de esa resolución se analizan con más detalle en la sección A.4 de la parte III del presente estudio.

En resumen, las constataciones empíricas complican ambas posturas extremas: el modelo entrenado no es, por regla general, una "copia" de su corpus de entrenamiento, pero tampoco puede considerarse desvinculado de dicho corpus. El alcance exacto variará según el modelo.

La segunda vía de análisis técnico es la bibliografía más amplia sobre los ataques de extracción de datos, que documenta situaciones en las que se pueden recuperar ejemplos individuales de entrenamiento concretos de un modelo entrenado mediante la formulación de *prompts* específicos.[23] Entre los ejemplos recuperados se incluyen fragmentos de texto de considerable extensión, imágenes identificables e información de carácter personal. Este trabajo se ha citado como prueba de que la extractabilidad puede ser una propiedad estructural de los grandes modelos generativos, y no solo el resultado de la manipulación por parte del usuario. Algunas empresas de IA sostienen que estas técnicas de formulación de *prompts* constituyen una manipulación indebida o una infracción de las condiciones de uso.[24] Sin embargo, las evaluaciones de este tipo se han vuelto más complejas durante el último año: los sistemas desplegados interponen cada vez más filtros y otras salvaguardias entre el modelo y el usuario para reducir la extractabilidad del contenido memorizado; en algunos estudios recientes se ha tenido que recurrir a técnicas específicas para forzar el funcionamiento de algunos sistemas, mientras que en otros se ha logrado obtener contenido protegido sin necesidad de ellas.[25]

---

[21] *The New York Times Company v. Microsoft Corp.*, n.º 1:23-cv-11195, escrito en apoyo de la solicitud de desestimación presentada por los demandados (Distrito Sur de Nueva York, 26 de febrero de 2024); Opinión y resolución (Distrito Sur de Nueva York, 4 de abril de 2025) (juez Stein) (por la que se deniega, en la medida pertinente, la solicitud de desestimación de las reclamaciones por infracción directa e indirecta). Para conocer la versión pública de OpenAI sobre este asunto, véase OpenAI, "Reporting the Facts About the New York Times' Lawsuit", disponible en openai.com/new-york-times.

[22] En *OpenAI, Inc. Copyright Infringement Litigation*, n.º 1:25-md-03143-SHS-OTW (Distrito Sur de Nueva York), opinión y resolución (juez Stein, 5 de enero de 2026), por la que se confirma la opinión y resolución (jueza Wang, 2 de diciembre de 2025) (por la que se desestima la solicitud de reexamen de OpenAI) y el auto subyacente de práctica de la prueba de noviembre de 2025. El tribunal ordenó la aportación de una muestra aleatorizada y anonimizada de 20 millones de registros de interacciones de usuarios de ChatGPT, extraída de la población de registros conservados, al considerar pertinente contar con un conjunto amplio de resultados generados para los usuarios a efectos del análisis del uso legítimo y, en particular, para determinar si los resultados del modelo funcionan como sustitutos en el mercado de las obras de los demandantes.

[23] Entre los trabajos anteriores en esta misma línea se encuentra el de Antonia Karamolegkou et al., "Copyright Violations and Large Language Models", arXiv:2310.13771 (2023) (en el que se demuestra, en distintos modelos de lenguaje, que es posible recuperar ejemplos concretos de los datos de entrenamiento mediante ataques de extracción).

[24] Véase Nicholas Carlini et al., "Extracting Training Data from Large Language Models", 30.º Simposio de Seguridad de USENIX (2021), pág. 2633; Milad Nasr et al., "Scalable Extraction of Training Data from Aligned, Production Language Models", Conferencia Internacional sobre Representaciones para el Aprendizaje (ICLR) (2025) (ampliación de Nasr et al., arXiv:2311.17035 (2023)).

[25] Ahmed Ahmed et al., "Extracting Books from Production Language Models", arXiv:2601.02671 (2026) (en el que se constató que, en algunos sistemas puestos en funcionamiento, era necesario recurrir a *prompts* formulados de

[Sigue la nota en la página siguiente]

Una analogía con los programas informáticos puede resultar útil para comprender la situación.[26] Cuando el código fuente de un programa (a menudo escrito en un lenguaje parecido al lenguaje natural) se compila en formato ejecutable, el resultado es diferente del código fuente y no es (fácilmente) legible para un ser humano. Sin embargo, la compilación no está desvinculada del original: el ejecutable conserva la información del programa fuente y, mediante ingeniería inversa, es posible reconstruir el original en gran medida. En ese sentido, el código ejecutable sigue siendo similar al programa fuente. Los pasos que permiten obtener un modelo entrenado a partir de un corpus de entrenamiento son, si seguimos esta analogía, posiblemente de naturaleza similar. En ambos casos, una fuente inteligible (un programa en código fuente; un corpus de obras) se transforma en un objeto numérico funcional que ya no se parece a ella, pero que sigue conteniendo información derivada de ella. En cualquier caso, esta analogía tiene sus límites: la compilación es un proceso determinístico que traslada la totalidad del programa fuente al ejecutable, mientras que la codificación de los datos de entrenamiento en un modelo es selectiva y, en parte, probabilística, por lo que la analogía es válida, como mucho, para el subconjunto de obras cuya memorización ha quedado demostrada; estudios recientes sostienen que es precisamente esta diferencia entre la recuperabilidad determinística y la meramente probabilística la que probablemente determinará la calificación jurídica del modelo.[27]

Las pruebas técnicas analizadas en este documento indican que un modelo entrenado conserva la información derivada de las obras incluidas en sus datos de entrenamiento, aunque la transformación sea considerable y el acceso a cualquier obra concreta pueda ser parcial y probabilístico. La forma en que esta observación técnica se traduce en la calificación jurídica del modelo y en la cuestión de si el propio modelo constituye una copia o una adaptación u obra derivada de las obras utilizadas para el entrenamiento, a efectos de la legislación de derecho de autor, es objeto de análisis tanto en los tribunales como en la bibliografía académica, como se expondrá más adelante en el presente estudio.

En este punto cabe señalar dos observaciones. En primer lugar, la memorización se utiliza como argumento tanto en la fase de entrenamiento, en la que se usa como prueba de que el entrenamiento del modelo implicó la codificación de copias, como en la fase de generación de resultados, como prueba de que el modelo desplegado puede generar obras protegidas en una forma que sustituye a las originales. En segundo lugar, las pruebas en las que se basan los demandantes se refieren a obras específicas en modelos concretos. Estas tienden a confirmar que la memorización puede ser considerable respecto de obras específicas y concretas que aparecen con frecuencia, pero no determinan que se trate de una propiedad uniforme de todos

manera artificiosa para obtener fragmentos de libros que el modelo había memorizado, mientras que en otros no era necesario hacerlo).

[26] Véase The Heart of the Matter, nota 10 supra, en la parte II.C (donde se establece una analogía entre el entrenamiento de un modelo de lenguaje y la compilación de un programa informático en formato ejecutable, y se señala que, en ambos casos, la obra original protegida por derecho de autor se traduce a otro formato que no es directamente comprensible por los seres humanos, pero en el que el contenido de la obra original no está ausente de esa versión comprensible para las máquinas).

[27] M. A. Lemley y A. F. Cooper, "Probabilistic 'Copies' in Generative AI Models", Berkeley Technology Law Journal (2026), disponible en https://ssrn.com/abstract=7067878 (en el que sostienen, partiendo de la bibliografía sobre extracción empírica analizada en esta parte, que solo debe considerarse que un modelo contiene una "copia" de una obra concreta cuando dicha obra pueda extraerse de forma sencilla y, en los casos más contundentes, de manera determinista; que, cuando la extracción solo se logra de forma probabilística, tras numerosos intentos, lo que se almacena no es una copia de la obra, sino un conjunto de "ingredientes"; y que, en consecuencia, se trata de una cuestión probatoria que debe responderse obra por obra y modelo por modelo).

los modelos ni que los parámetros de un modelo se reduzcan, en general, a copias almacenadas. La calificación jurídica de estas cuestiones se aborda en la parte III.[28]

Por último, también cabe señalar que un modelo puede cambiar una vez finalizado el entrenamiento mediante el ajuste fino, el entrenamiento de alineación o los intentos de "desaprendizaje automático" dirigidos a contenidos específicos. El hecho de que un modelo deje de reproducir una obra determinada tras una intervención de este tipo no permite concluir que dicha obra nunca hubiera sido codificada en él. La bibliografía técnica señala que sigue siendo difícil eliminar de forma fiable obras concretas de un modelo entrenado y que este tipo de intervenciones pueden deteriorar el modelo de formas imprevistas.[29]

3. La relación entre los datos de entrada y los de salida

Una vez desplegado, el modelo recibe datos de entrada del usuario (un *prompt*) y genera unos datos de salida (un resultado). El resultado puede ser texto, código, imágenes, audio, video u otro tipo de contenido. La relación entre los datos de entrada y los de salida es, en algunos aspectos, análoga a la relación entre una consulta y una respuesta en un motor de búsqueda, pero la analogía es parcial. El modelo no consulta un corpus indexado ni extrae fragmentos de él, sino que genera contenido nuevo mediante el muestreo de la distribución de probabilidad que ha interiorizado durante el entrenamiento. En algunos casos, lo que genera está estrechamente relacionado con obras concretas de sus datos de entrenamiento; en otros, el resultado es una composición novedosa que no se asemeja mucho a ningún ejemplo concreto del entrenamiento.

Hay tres observaciones que resultan especialmente relevantes. En primer lugar, el resultado puede ser una reproducción literal o casi literal de una obra incluida en los datos de entrenamiento, en cuyo caso constituye una reproducción de la obra original.[30] Así se ha demostrado en estudios sobre modelos de lenguaje que reproducen fragmentos de libros protegidos por derechos de autor y sobre modelos de generación de imágenes que reproducen fotografías y obras de arte protegidas.[31] Que dicha reproducción se considere una infracción dependerá de una serie de cuestiones adicionales, entre las que se incluyen la aplicabilidad de excepciones y limitaciones, la función de los filtros de salida incorporados por el desarrollador y la atribución de responsabilidad entre el desarrollador del modelo y el usuario que haya

[28] En cuanto a los procedimientos consolidados, véase *In re OpenAI, Inc. Copyright Infringement Litigation*, n.º 1:25-md-03143-SHS-OTW (Distrito Sur de Nueva York), en el que se han consolidado las demandas interpuestas por *The New York Times* y diversos grupos de autores contra OpenAI y Microsoft. La caracterización del modelo entrenado como un mecanismo de almacenamiento y recuperación, y de los parámetros como una forma comprimida de las obras incorporadas, refleja las alegaciones de los demandantes; la caracterización contraria de los parámetros como generalizaciones estadísticas que codifican patrones lingüísticos refleja la postura de los demandados. En cuanto a la bibliografía empírica que advierte contra ambas caracterizaciones, véase el estudio de Cooper *et al*. citado anteriormente en esta sección.

[29] A. Feder Cooper *et al.*, "Machine Unlearning Doesn't Do What You Think: Lessons for Generative AI Policy, Research, and Practice", 39.ª Conferencia sobre Sistemas de Procesamiento de Información Neuronal (NeurIPS) (2025).

[30] Cuando el resultado generado por un gran modelo de lenguaje (LLM) es similar a una obra protegida ya existente que el modelo no utilizó para su entrenamiento y a la que no tuvo acceso de ningún otro modo, resulta mucho menos evidente que se haya realizado una "copia".

[31] La reproducción literal por ciertos modelos de lenguaje ha quedado documentada en estudios académicos y en escritos procesales; la reproducción de fotografías y obras de arte reconocibles por parte de ciertos modelos de generación de imágenes ha quedado documentada en relación con los modelos de difusión. El número de ejemplos documentados no influye directamente en la calificación jurídica de ningún resultado concreto, ya que esta depende de la legislación nacional aplicable y de los hechos del caso.

formulado el *prompt*. Esas cuestiones se mencionan brevemente más adelante en el presente estudio.[32]

En segundo lugar, el resultado puede ser sustancialmente similar a una obra incluida en los datos de entrenamiento sin que se trate de una reproducción literal. Esto podría constituir una infracción, por ejemplo, en virtud de la doctrina respecto a la similitud sustancial, una norma habitual en varias legislaciones nacionales en materia de derecho de autor.

En tercer lugar, el resultado puede ser totalmente nuevo, en el sentido de que no se pueda identificar ninguna obra concreta de los datos de entrenamiento como su fuente. Un modelo generativo es capaz de generar contenido que no se corresponde con ningún ejemplo de entrenamiento concreto. De hecho, en la mayoría de los casos eso es precisamente lo que hace. En tales casos, la relación entre la obra y cualquier obra protegida concreta se ve atenuada. La cuestión de si el resultado sigue “basándose” en los datos de entrenamiento, en cualquier sentido que la legislación de derecho de autor pueda reconocer, es una cuestión que se ha planteado ante los tribunales de varias jurisdicciones, pero sobre la que aún no se ha alcanzado un consenso.

Por lo tanto, en la parte I se identifica el uso de obras protegidas y cuestiones relativas a su conservación o reproducción en las distintas fases del desarrollo y funcionamiento de un modelo de IA generativa: en el corpus que se recopila al inicio, en las operaciones de almacenamiento y procesamiento que se aplican al corpus durante el entrenamiento, en el propio modelo entrenado y en los datos de salida generados tras el despliegue. En cada fase pueden entrar en juego los principales derechos exclusivos reconocidos por las legislaciones nacionales y los tratados internacionales en materia de derecho de autor. En las siguientes partes de este estudio se abordan la cuestión de si cada reproducción puede constituir una infracción *prima facie* (salvo que se aplique una excepción o limitación) y la cuestión de qué mecanismos existen o podrían desarrollarse para autorizar mediante licencia los usos pertinentes.

PARTE II. AUTORIZACIÓN, CONCESIÓN DE LICENCIAS Y COMPENSACIÓN

La parte I de este estudio analiza, desde una perspectiva técnica, en qué fases del desarrollo y del funcionamiento de un modelo de IA generativa pueden verse afectados los principales derechos exclusivos que reconocen las leyes nacionales de derecho de autor y los tratados internacionales. La parte que ahora nos ocupa pasa de dichos derechos a los mecanismos mediante los cuales pueden ejercerse, ser objeto de renuncia o ser sustituidos en el contexto de la IA. La cuestión de si los usos pertinentes son lícitos sin autorización y, cuando no lo son, la cuestión de cómo pueden los desarrolladores dotar a sus operaciones de una base lícita y de cómo debe compensarse a los titulares de los derechos en ese proceso, conducen a un único corpus doctrinal: el derecho relativo a la concesión de licencias de derecho de autor, individuales y colectivas, voluntarias y legales, junto con las excepciones y limitaciones que condicionan la necesidad misma de obtener una licencia.

La presente parte se estructura de la siguiente manera. En la sección A se examinan los principales cauces a través de los cuales los desarrolladores de IA adquieren actualmente la materia susceptible de protección por derecho de autor con la que entrenan sus modelos, así como la situación jurídica de cada uno de ellos. En la sección B se abordan las principales cuestiones de política que configuran actualmente las prácticas de concesión de licencias: los derechos afectados por el entrenamiento, la concesión de licencias voluntarias, la concesión de

[32] En cuanto a la relación entre el *prompt* introducido por el usuario, el modelo y el resultado generado, y a la cuestión aún sin resolver de cómo debe repartirse la responsabilidad entre el desarrollador y el usuario, véase la parte III *infra*.

licencias colectivas en sus tres formas principales, la situación de la excepción relativa a la minería de textos y datos de la UE y el debate académico sobre su pertinencia para el entrenamiento de la IA, y los regímenes de remuneración. Se mencionan distintos países como ejemplos de los planteamientos adoptados en distintas partes del mundo; no se realiza un análisis en profundidad ni se presenta un panorama exhaustivo.

SECCIÓN A. OPCIONES ACTUALES DE LOS DESARROLLADORES PARA OBTENER MATERIA SUSCEPTIBLE DE PROTECCIÓN POR DERECHO DE AUTOR CON FINES DE ENTRENAMIENTO

1. Las cinco vías de adquisición

En la práctica, los desarrolladores de IA adquieren sus datos de entrenamiento por una de las siguientes cinco vías, o mediante una combinación de ellas. Una de ellas es la utilización de conjuntos de datos abiertos o seleccionados creados por terceros que a veces adquieren de forma gratuita o tras el pago de una tasa. Un ejemplo destacado es Common Crawl, un repositorio de páginas web que se actualiza continuamente, que una fundación sin ánimo de lucro rastrea y facilita, y que sustenta, directa o indirectamente, los corpus de entrenamiento textuales de la mayoría de los principales modelos de lenguaje. LAION-5B y sus predecesores son el equivalente para los modelos de generación de imágenes.[33] Esos conjuntos de datos son a su vez el resultado de operaciones de recuperación automatizada a gran escala y pueden arrastrar las cuestiones de derecho de autor vinculadas a la recuperación subyacente. La situación jurídica dependerá de la base sobre la que se haya obtenido el material de origen y de la legislación aplicable a los usos posteriores del conjunto de datos. Ha empezado a surgir, en el sector público, un equivalente a estos conjuntos de datos seleccionados: en la República de Corea se ha actualizado el sistema de licencias de datos públicos (Gonggong Nuri) para incluir una categoría de licencias en virtud de la cual los conjuntos de datos públicos llevan una indicación expresa de que se autoriza su uso para el entrenamiento de modelos de IA.[34]

Otra fuente son las denominadas “bibliotecas en la sombra”, grandes repositorios no autorizados de libros, artículos, música y otras obras, que, según se informa o se alega, han sido utilizados por varios modelos de uso extendido, ya sea directamente o a través de conjuntos de datos que los incluyen. Las conclusiones de los tribunales de los Estados Unidos de América sobre piratería que se examinan en la parte III, sección A, más adelante, ilustran el riesgo jurídico que conlleva recurrir a ese tipo de repositorios.

A efectos de comparación, algunos tribunales han distinguido entre la adquisición ilícita de obras y una cuestión distinta, a saber, si el uso de las obras para el entrenamiento está amparado por una excepción o limitación. En una de las resoluciones estadounidenses que se examinan más adelante, se determinó que la descarga y conservación de copias pirateadas no constituía uso leal, mientras que el entrenamiento de un modelo con copias de las mismas obras adquiridas lícitamente se consideró transformador y amparado por dicho uso leal. En

[33] Más información sobre Common Crawl en https://commoncrawl.org. El conjunto de datos LAION-5B (utilizado en la creación de Stable Diffusion y de otros modelos de generación de imágenes) y su predecesor, LAION-400M, son conjuntos de datos de imágenes y textos disponibles públicamente recopilados por la red Large-scale Artificial Intelligence Open Network. El conjunto de datos en sí no contiene las imágenes originales, sino pares de direcciones URL y descripciones textuales; las imágenes se recuperan de sus ubicaciones originales en el momento del entrenamiento. Sobre la calificación jurídica de las operaciones de compilación de conjuntos de datos como minería de textos y datos, véanse *Kneschke v. LAION e.V.*, Tribunal Regional de Hamburgo, 27 de septiembre de 2024 (310 O 227/23) y, en apelación, *Kneschke v. LAION e.V.*, Tribunal Regional Superior de Hamburgo, 10 de diciembre de 2025 (5 U 104/24).

[34] Se hace referencia al sistema Gonggong Nuri (Licencia de Gobierno Abierto de Corea), en virtud del cual las instituciones públicas ponen a disposición datos públicos que pueden ser reutilizados con arreglo a categorías normalizadas de licencia; el marco actualizado añade una categoría que permite expresamente su uso para el entrenamiento de modelos de IA.

este sentido, el carácter lícito de la adquisición y el del entrenamiento son independientes desde el punto de vista analítico. La distinción es importante porque, en un mismo modelo, los datos de entrenamiento pueden proceder de fuentes lícitas e ilícitas, lo que añade complejidad a la concesión de licencias y la observancia. Por lo tanto, es posible que un desarrollador necesite tanto adquirir lícitamente el corpus como contar con una base lícita para el propio entrenamiento, con sujeción a la legislación o legislaciones nacionales aplicables.

La tercera vía es la concesión de licencias individuales y colectivas. Un desarrollador puede adquirir el acceso a un corpus seleccionado o negociarlo con una editorial, una organización de noticias, un organismo de radiodifusión, un banco de imágenes, un operador de bases de datos académicas o cualquiera de los titulares institucionales de derechos. En los últimos tres años, la concesión de licencias se ha convertido en una práctica habitual, y los principales desarrolladores de IA han firmado acuerdos con grandes organizaciones de noticias y bancos de imágenes de Europa, América y Asia.[35] El rasgo estructural de la concesión de licencias es que el desarrollador obtiene una cadena de titularidad fehaciente sobre el corpus objeto de licencia, al menos en lo que respecta a los derechos propios del licenciante. Las limitaciones de esa cadena se examinan más adelante, en la sección B.2. En algunos casos, el contrato incluye la transmisión del contenido.

La cuarta vía, que es una de las principales en lo que respecta al volumen, es la recopilación directa de material de la web abierta mediante el rastreo y el raspado web que realiza el propio desarrollador, y no un tercero. Los procedimientos técnicos se describen en la sección A.2 de la parte I del presente estudio. La magnitud de la recuperación automatizada es considerable: según algunas estimaciones, los agentes automatizados ya representan la mayor parte del tráfico web mundial.[36] El rastreo suele realizarse sin el consentimiento previo de los titulares de los derechos sobre el material recuperado, según una de estas tres teorías jurídicas: que el material es de dominio público o ha sido objeto de una licencia válida al público en condiciones permisivas; que la operación está amparada por una excepción o limitación de la legislación de derecho de autor aplicable (en la Unión Europea, la excepción relativa a la minería de textos y datos del artículo 4 de la Directiva sobre los derechos de autor en el mercado único digital; en los Estados Unidos de América y en la República de Corea, en este último caso en virtud del artículo 35-5 de la Ley de Derecho de Autor, una doctrina general del uso leal; en Singapur y el Japón, disposiciones legales específicas); o que la operación no afecta en absoluto al derecho de reproducción porque las copias realizadas son provisionales o transitorias o, en algunas jurisdicciones, porque dichas copias están amparadas por una excepción expresa para los actos de reproducción provisional (como en el caso del artículo 5.1 de la Directiva InfoSoc de la Unión Europea). Estas teorías no se excluyen entre sí, y un desarrollador puede invocar una u otra según la parte del corpus de que se trate.

La quinta vía son los datos sintéticos: datos generados por un modelo de IA existente y utilizados para entrenar otro o para realizar ajustes finos en él. Los datos sintéticos son una herramienta cada vez más utilizada, sobre todo en las fases posteriores del desarrollo de modelos, en particular al realizar un ajuste fino supervisado y en el modelado de

[35] En relación con los acuerdos bilaterales de licencia entre desarrolladores de IA y editoriales, véanse (entre otros muchos) los acuerdos firmados por OpenAI con Associated Press, News Corp, Le Monde, Prisa Media, Axel Springer, Vox Media, The Atlantic y Hearst; por Google con Reddit y Associated Press; y por Anthropic con diversas editoriales. Estos acuerdos se han divulgado con distintos grados de detalle y sus condiciones financieras son confidenciales por motivos comerciales.

[36] Véase Imperva, *2025 Bad Bot Report* (duodécima edición, 2025), donde se indica que el tráfico automatizado representó el 51 % de todo el tráfico de Internet en 2024 y superó al tráfico humano por primera vez en una década, y que los denominados “bots maliciosos” representaron el 37 %.

recompensas.[37] Desde el punto de vista del derecho de autor, los datos sintéticos plantean una cuestión distinta. Si el modelo que los generó fue entrenado con obras protegidas, ¿los productos sintéticos también reproducen o adaptan la expresión protegida? Esta cuestión aún no se ha planteado ante un tribunal de forma que permita obtener una respuesta vinculante.

2. Acuerdos de colaboración

Por último, además de estas cinco grandes vías, en algunas jurisdicciones se observan formas de colaboración organizada entre desarrolladores de IA y titulares de derechos institucionales o de propiedad estatal, especialmente pertinentes para determinadas aplicaciones sectoriales. China ofrece un ejemplo. El desarrollo de grandes modelos especializados en ámbitos verticales ha dado lugar a colaboraciones entre un desarrollador de modelos de IA de uso general y emisoras de radio y televisión de Hong Kong (para datos de voz en cantonés), entre desarrolladores y editoriales en línea con amplios repertorios de escritores contratados (para novelas digitales), entre desarrolladores y grandes repositorios de imágenes con miles de millones de imágenes adquiridas mediante cesión o licencia exclusiva (para datos de entrenamiento visuales), y entre desarrolladores y operadores de bases de datos académicas que construyen conjuntamente modelos especializados. Se afirma que esta práctica está muy extendida, aunque los propios acuerdos no suelen denominarse licencias de derecho de autor y no siempre se revela su alcance exacto. Como se expone más adelante, en la sección B.2, el límite estructural de este planteamiento es que la editorial, el organismo de radiodifusión o el repositorio suelen ser titulares únicamente de determinados derechos concedidos por el autor de la obra original, y no está claro que los derechos de los que son titulares incluyan el entrenamiento de la IA.

SECCIÓN B. PRINCIPALES ELEMENTOS DE POLÍTICA

1. Derechos afectados por el entrenamiento de los modelos de IA generativa

a) Derecho de reproducción

Las operaciones técnicas descritas en la parte I afectan, a primera vista, a varios de los derechos exclusivos reconocidos por las legislaciones nacionales de derecho de autor y por los tratados internacionales en la materia. El derecho de reproducción ocupa un lugar fundamental en este análisis.[38] La constitución de un corpus de entrenamiento requiere realizar copias; el almacenamiento y el tratamiento de dicho corpus durante las distintas fases del entrenamiento implican nuevas copias, algunas persistentes y otras transitorias; y determinados aspectos del modelo entrenado, según el análisis técnico expuesto en la parte I, sección C.2, pueden analizarse a su vez como copias de las obras con las que se entrenó el modelo. Cada una de estas operaciones se enmarca, a primera vista, en el derecho de reproducción reconocido en el artículo 9.1) del Convenio de Berna, en el artículo 1.4) del Tratado de la OMPI sobre Derecho de Autor, en el artículo 7 del WPPT y del Tratado de Beijing y en las disposiciones correspondientes de las legislaciones nacionales.

El tratamiento de las copias temporales y transitorias requiere un examen separado, porque dos técnicas jurídicas distintas conducen a resultados prácticos similares: en algunas jurisdicciones se considera que esas copias no forman parte del ámbito del derecho de reproducción, mientras que en otras se les da cabida mediante una excepción expresa. China ofrece el principal ejemplo de la primera técnica: se ha considerado de forma sistemática que la reproducción temporal queda fuera del ámbito del derecho. Así, en la práctica judicial china, las

[37] Los datos sintéticos son datos generados por un modelo de IA existente y utilizados en el entrenamiento o el ajuste fino de otro modelo. Plantean cuestiones específicas de derecho de autor cuando el propio modelo generador se entrenó con obras protegidas. Sobre su uso creciente en las fases posteriores del desarrollo de modelos, véase R. Liu et al., "Best Practices and Lessons Learned on Synthetic Data", arXiv:2404.07503 (2024), presentado en la Primera Conferencia sobre Modelado del Lenguaje.

[38] Véase la parte I, secciones A.1 y A.2.

reproducciones temporales excluidas son las copias transitorias realizadas en una memoria de acceso aleatorio o una memoria caché que son estrictamente necesarias para un proceso de tratamiento de datos; no existe una posición consolidada sobre el tiempo exacto que puede subsistir una copia para seguir considerándose temporal. El artículo 10.1.5 de la Ley de Derecho de Autor define la reproducción como el derecho a realizar una o varias copias de una obra mediante procedimientos como la impresión, la fotocopia, el grabado, la grabación sonora, la grabación en video, la duplicación, la fotografía o la digitalización,[39] y la delegación china, entre otras, votó en contra de la declaración concertada de la Conferencia Diplomática que habría incluido claramente el almacenamiento digital temporal en el ámbito del derecho.[40] El Reglamento sobre la Protección del Derecho de Comunicación al Público a través de Redes de Información (2006) no incluyó ninguna disposición que reconociera la reproducción provisional como acto de reproducción, y los tribunales chinos han aplicado este criterio en casos relacionados con la transcodificación y operaciones digitales transitorias similares. Desde esta perspectiva, es posible que el acto de entrenamiento en sí, en la medida en que solo implique un acceso transitorio a obras en la memoria o en la memoria caché que desaparece al finalizar la ejecución del entrenamiento, no afecte al derecho de reproducción contemplado en la legislación china, aunque la constitución de la base de datos subyacente de obras indudablemente sí podría. En otras jurisdicciones, se llega al mismo resultado práctico mediante una excepción expresa: en la Unión Europea, el artículo 5, apartado 1, de la Directiva InfoSoc admite las copias transitorias y accesorias mediante una excepción de alcance limitado, en lugar de excluirlas del derecho *ab initio*. Las copias en cuestión son, en esencia, las mismas en ambos sistemas; la diferencia radica en la técnica jurídica empleada.[41] En la mayoría de las demás jurisdicciones examinadas, se considera en general que el acto de entrenamiento afecta al derecho de reproducción, y la cuestión analítica depende de que exista una excepción o limitación aplicable.

b) El derecho de adaptación y las obras derivadas

Además del derecho de reproducción, hay otros tres derechos que se ven o pueden verse afectados. El derecho de adaptación, reconocido en el artículo 12 del Convenio de Berna y en las disposiciones nacionales correspondientes, se invoca en la fase de generación de resultados, ya que se considera que la generación de resultados supone una transformación o alteración de las obras originales. Por ejemplo, la legislación brasileña incluye expresamente el derecho de adaptación entre las formas de utilización que requieren autorización previa y expresa,[42] y la Secretaría de Derecho de Autor y Propiedad Intelectual del Ministerio de Cultura ha adoptado la posición de que la generación de productos afecta al derecho de adaptación, además de al derecho de reproducción. Las legislaciones de Alemania, el Canadá y Francia contemplan posicionamientos similares, en líneas generales.

---

[39] Ley de Derecho de Autor de la República Popular China, artículo 10.1.5.

[40] Véase el documento de la OMPI CRNR/DC/101, Actas resumidas, sesión plenaria, párrafos 497 a 502. No obstante, la Conferencia Diplomática adoptó la Declaración concertada respecto del artículo 1.4) del WCT; China votó en contra.

[41] Artículo 5, apartado 1, de la 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 los derechos de autor en la sociedad de la información, DO L 167/10 [a menudo denominada “Directiva InfoSoc”]. La excepción abarca los actos de reproducción provisional que sean transitorios o accesorios y formen parte integrante y esencial de un proceso tecnológico, cuya única finalidad consista en facilitar una transmisión en una red entre terceras partes por un intermediario o una utilización lícita, y que no tengan por sí mismos una significación económica independiente.

[42] Ley de Derecho de Autor del Brasil (Ley n.º 9610/1998), artículo 29, puntos III y IV (adaptación, arreglos musicales, traducciones y otras transformaciones). El Ministerio de Cultura adoptó esa posición, entre otras cosas, en una nota técnica dirigida a la Autoridade Nacional de Proteção de Dados en octubre de 2024 relativa al entrenamiento de los modelos de un gran desarrollador, en la que invocaba el artículo 29.III (transformación para la realización de obras derivadas); véase la información publicada en https://nucleo.jor.br/curtas/2024/10/10/ministerio-cultura-prejuizos-direitos-ia-meta/, consultada el 7 de julio de 2026.

c) Comunicación al público y puesta a disposición

El derecho de puesta a disposición del público, previsto en el artículo 8 del Tratado de la OMPI sobre Derecho de Autor y en los artículos 10 y 14 del Tratado de la OMPI sobre Interpretación o Ejecución y Fonogramas, entra en juego principalmente en la fase de generación de resultados, cuando el modelo desplegado devuelve contenido al usuario, si dicho contenido incluye una copia de la expresión protegida. En el Brasil, país que no es Parte Contratante del WCT ni del WPPT, el *Superior Tribunal de Justiça* resolvió en 2017, en el REsp 1.559.264-RJ, que la transmisión por flujo continuo de música por Internet afecta al derecho de comunicación al público, que abarca lo que otras jurisdicciones podrían identificar como el derecho de puesta a disposición.[43] Esto está en consonancia con la "solución marco" adoptada en la Conferencia Diplomática de 1996, en virtud de la cual el derecho de puesta a disposición puede aplicarse mediante el derecho de comunicación al público, el derecho de distribución o una combinación de ambos. Los tribunales brasileños todavía no se han pronunciado sobre si este mismo razonamiento puede aplicarse a los resultados que genera un modelo de IA ya desplegado.[44] Los artistas intérpretes o ejecutantes y los productores de fonogramas son titulares de derechos paralelos de puesta a disposición en virtud de los artículos 10 y 14 del WPPT anteriormente mencionado, por lo que el análisis va más allá de los derechos de los autores.

d) Derechos morales

El derecho moral también puede verse afectado. El derecho a la integridad es el que se invoca con más frecuencia. En el Brasil, se considera que la fragmentación de las obras durante el proceso de entrenamiento puede vulnerar el derecho a la integridad, y que no mencionar a los autores de los materiales utilizados puede vulnerar el derecho de paternidad.[45] La Ley de Derecho de Autor del Canadá reconoce tanto el derecho a la integridad como el derecho de paternidad como derechos morales del autor.[46] En varias jurisdicciones del sistema de *droit d'auteur*, se ha planteado la cuestión de si el uso de una obra en un corpus de entrenamiento sin mención de su autoría, y de una manera que puede generar resultados que se perciban como versiones despectivas o deformadas del original, constituye una vulneración de los derechos morales en la doctrina académica, pero todavía no se ha recibido una respuesta autorizada.

Por último, en la Unión Europea y en algunas jurisdicciones que han adoptado una protección similar, puede verse afectado el derecho sui géneris sobre las bases de datos protegido por la Directiva 96/9/CE cuando el corpus de entrenamiento se constituye mediante la extracción y reutilización de partes sustanciales de una base de datos protegida.[47] Este derecho es independiente del derecho de autor y tiene sus propios criterios de inversión sustancial y de "parte sustancial" (en el caso del "derecho de extracción"). El alcance de su aplicación a la constitución de corpus de entrenamiento de IA sigue sin estar resuelto, también en relación con las bases de datos de información jurídica, los agregadores de revistas científicas y los grandes bancos de imágenes.

---

[43] Superior Tribunal de Justiça, REsp 1.559.264-RJ (2017) (en el que se resuelve que la transmisión por flujo continuo de música por Internet constituye una forma de ejecución pública y afecta a los derechos de comunicación al público, y que, sobre esa base, el ECAD está facultado para recaudar).

[44] Ley de Derecho de Autor del Brasil (Ley n.º 9.610/1998), artículo 5.V) (definición de "comunicación al público") y artículo 29.VII (distribución mediante cualquier sistema que permita al usuario seleccionar la obra).

[45] Ley de Derecho de Autor del Brasil (Ley n.º 9610/1998), artículo 24, puntos I a IV. Los artistas intérpretes o ejecutantes también gozan de los derechos morales de integridad y paternidad sobre sus interpretaciones o ejecuciones en virtud del artículo 92.

[46] Ley de Derecho de Autor, R.S.C. 1985, c. C-42, arts. 14.1 y 28.1.

[47] Directiva 96/9/CE del Parlamento Europeo y del Consejo, de 11 de marzo de 1996, sobre la protección jurídica de las bases de datos, DO L 77/20, artículo 7 (el derecho sui géneris). Véase también TJUE, C-203/02, *British Horseracing Board y William Hill*, ECLI:EU:C:2004:695, sobre las condiciones del criterio de inversión sustancial.

Por consiguiente, los derechos que pueden verse afectados por el entrenamiento de un modelo de IA generativa pueden ir más allá del derecho de reproducción. Por tanto, el que una licencia bilateral o colectiva cubra los usos pertinentes depende de los derechos que incluya; una licencia limitada a la reproducción, por ejemplo, puede no proteger frente a reclamaciones basadas en la adaptación, la puesta a disposición, los derechos morales o, cuando proceda, el derecho sui géneris sobre las bases de datos. Otra cuestión, que se retoma en la parte III en el contexto de la observancia transfronteriza, es que el conjunto de derechos y el alcance exacto de cada uno de ellos varían de una jurisdicción a otra, de modo que un acuerdo de licencia concebido para obtener los derechos en una jurisdicción puede dejar al desarrollador expuesto en otra.[48]

e) Información sobre la gestión de derechos y medidas contra la elusión

En virtud de los artículos 11 y 12 del WCT, los artículos 18 y 19 del WPPT y los artículos 15 y 16 del Tratado de Beijing, los titulares de derechos tienen derecho a una protección jurídica adecuada y a recursos legales efectivos contra la elusión de las medidas tecnológicas efectivas que utilicen en relación con el ejercicio de sus derechos en virtud de dichos instrumentos y del Convenio de Berna, así como a medidas de reparación adecuadas y eficaces contra cualquier persona que realice conscientemente los siguientes actos, sabiendo —o, a efectos de los recursos civiles, teniendo motivos razonables para saber— que ello inducirá, permitirá, facilitará u ocultará una infracción de cualquier derecho amparado por dichos Tratados o por el Convenio de Berna:

i) la eliminación o alteración sin autorización de cualquier información electrónica sobre la gestión de derechos;

ii) la distribución, la importación con fines de distribución, la radiodifusión o la comunicación al público, sin autorización, de obras o copias de obras, fonogramas o interpretaciones y (conforme a los instrumentos respectivos) a sabiendas de que se ha suprimido o alterado sin autorización la información sobre la gestión electrónica de derechos.

Según nuestra información, en el momento de redactar este estudio no se había interpuesto ninguna demanda con resultado favorable contra empresas en relación con el entrenamiento de modelos de IA.

2. Concesión voluntaria de licencias (colectivas e individuales)

a) Concesión de licencias individuales y colectivas para el uso de la IA

La concesión de licencias voluntarias implica la negociación de una autorización entre un desarrollador de IA y un titular de derechos o un representante de múltiples titulares de derechos, como un OGC u otro intermediario, sin que medie ningún mandato legal. Este es el régimen por defecto en todas las jurisdicciones estudiadas.

En la práctica, la concesión de licencias voluntarias se caracteriza por dos rasgos estructurales. El primero es la asimetría entre el reducido número de grandes titulares institucionales de derechos que pueden negociar con los desarrolladores en condiciones comercialmente significativas y el elevado número de autores individuales y pequeños titulares de derechos para quienes esa negociación suele resultar inviable debido a los costos de transacción, la falta de conocimientos especializados u otros factores. Los grandes acuerdos bilaterales celebrados en los últimos tres años se han concluido entre desarrolladores de IA y grandes editoriales,

[48] Véase *Toronto Star Newspapers Ltd v. OpenAI Inc.*, 2025, Tribunal Superior de Justicia de Ontario, 6217 (juez Kimmel., 7 de noviembre de 2025) (en el que se rechaza la excepción de falta de jurisdicción de OpenAI por existir una conexión real y sustancial con Ontario).

organizaciones de noticias, organismos de radiodifusión y bancos de imágenes; no existe una infraestructura comparable para los autores, fotógrafos, ilustradores o músicos individuales, que a menudo dependen de los acuerdos celebrados por sus editoriales y otros licenciatarios para obtener una parte del valor generado.

Algunos de estos acuerdos han alcanzado ya una escala considerable. Se ha informado de que una gran editorial de prensa y un desarrollador de IA han firmado un acuerdo de cinco años que superó los 250 millones de dólares de los Estados Unidos,[49] y la concesión de licencias de contenidos para uso computacional se ha convertido en una fuente de ingresos importante para los grandes bancos de imágenes: en sus resultados del ejercicio completo de 2025, Shutterstock, Inc. declaró unos ingresos de 203,3 millones de dólares de los Estados Unidos en la categoría de datos, distribución y servicios, lo que representa aproximadamente el 20,5 % de unos ingresos totales de 989,9 millones de dólares de los Estados Unidos; por su parte, Getty Images declaró unos ingresos de 981,3 millones de dólares de los Estados Unidos en el ejercicio completo de 2025. El 7 de enero de 2025 las empresas anunciaron una fusión entre iguales; el 30 de junio de 2026, Getty Images anunció que pondría fin al acuerdo de fusión después de que la Autoridad de Competencia y Mercados del Reino Unido condicionara su autorización a la desinversión de la actividad editorial de Shutterstock. Según la información publicada, el acuerdo se rescindió el 7 de julio de 2026.[50] La selectividad de este mercado queda ilustrada por la prensa francesa: en marzo de 2024, Le Monde celebró un acuerdo con OpenAI que se describió como el primero entre un medio de comunicación francés y un gran desarrollador de IA generativa.[51] Ese mismo año, la APIG trató de entablar conversaciones colectivas con OpenAI, aunque la información disponible indica que por entonces dichas conversaciones no dieron lugar a un acuerdo de licencia más amplio.[52] Como respuesta a esta asimetría, han empezado a surgir intermediarios que agrupan los derechos de editoriales más pequeñas: según una fuente de información empresarial, varios miles de las editoriales que forman parte de la red de uno de estos intermediarios, que es mucho más amplia, han decidido

[49] *The Wall Street Journal* informó de que el acuerdo de cinco años entre News Corp y OpenAI superó los 250 millones de dólares de los Estados Unidos. Las condiciones económicas de este tipo de acuerdos suelen ser confidenciales, y es la prensa la que facilita las cifras, sin que las partes las hayan confirmado públicamente. Véase "OpenAI, WSJ Owner News Corp Strike Content Deal Valued at Over $250 Million" (The Wall Street Journal, 22 de mayo de 2024) <https://www.wsj.com/business/media/openai-news-corp-strike-deal-23f186ba>, consultado el 3 de julio de 2026; "Sam Altman's OpenAI signs content agreement with News Corp" (Reuters, 22 de mayo de 2024) <https://finance.yahoo.com/news/sam-altmans-openai-signs-content-201808782.html>, consultado el 3 de julio de 2026.

[50] Shutterstock, "Shutterstock Reports Full Year 2025 and Fourth Quarter Financial Results" (17 de febrero de 2026) <https://investor.shutterstock.com/node/14551/pdf>, consultado el 3 de julio de 2026; Getty Images Holdings, Inc, "Getty Images Reports Fourth Quarter and Full Year 2025 Results" (16 de marzo de 2026) <https://investors.gettyimages.com/node/15726/pdf>, consultado el 3 de julio de 2026; Getty Images Holdings, Inc y Shutterstock, Inc, "Getty Images and Shutterstock to Merge, Creating a Premier Visual Content Company" (7 de enero de 2025) <https://www.sec.gov/Archives/edgar/data/1898496/000110465925001732/tm252186d1_ex99-1.htm>, consultado el 3 de julio de 2026. Sobre la rescisión de la fusión, véase Competition and Markets Authority, "Getty abandons Shutterstock merger plans" (julio de 2026) <https://www.gov.uk/government/news/getty-abandons-shutterstock-merger-plans>, consultado el 31 de julio de 2026.

[51] *Le Monde* describió el acuerdo, firmado en marzo de 2024, como el primero entre un medio de comunicación francés y un gran desarrollador de modelos de IA generativa. Véase "Le Monde signs artificial intelligence partnership agreement with OpenAI" (*Le Monde*, 13 de marzo de 2024) <https://www.lemonde.fr/en/about-us/article/2024/03/13/le-monde-signs-artificial-intelligence-partnership-agreement-with-open-ai_6615418_115.html>, consultado el 3 de julio de 2026.

[52] Se trata de la Alliance de la presse d'information générale (APIG), que, en junio de 2024 y junto con el Syndicat des éditeurs de la presse magazine (SEPM), solicitó formalmente la apertura de negociaciones con OpenAI; según la información publicada, el desarrollador respondió que, en ese momento, no veía la necesidad de celebrar nuevos acuerdos de licencia con editoriales francesas. Véase "Intelligence artificielle : un accord de partenariat entre Le Monde et OpenAI" (Le Monde, 13 de marzo de 2024) <https://www.lemonde.fr/le-monde-et-vous/article/2024/03/13/intelligence-artificielle-un-accord-de-partenariat-entre-le-monde-et-openai_6221836_6065879.html>, consultado el 3 de julio de 2026.

conceder licencias sobre sus contenidos para usos de IA;[53] del mismo modo, el desarrollador de un motor de respuestas basado en contenidos objeto de licencia afirma que también se está agrupando a las publicaciones que participan en su servicio.[54]

El segundo rasgo estructural es la cuestión de la cadena de titularidad. En muchas jurisdicciones, cuando el autor de la obra original ha concedido determinados derechos a una editorial, un organismo de radiodifusión o un repositorio mediante un contrato de licencia o de cesión celebrado antes de la llegada de la IA generativa, la cuestión de si esos derechos se extienden al uso de la obra en el entrenamiento de un modelo de IA, además de ser jurídicamente incierta, no está contemplada en el contrato. Esta incertidumbre no es exclusiva de la IA: se plantean cuestiones comparables en los contratos entre autores y editores de música, entre artistas intérpretes o ejecutantes y productores de fonogramas, y entre productores audiovisuales y los artistas contratados para una producción, siempre que una modalidad de explotación no estuviera contemplada en el momento de la contratación; lo que distingue el contexto de la IA es la escala a la que se plantea la cuestión y la dificultad práctica de renegociar. En China, por ejemplo, la mayoría de los autores conceden a sus editoriales el derecho exclusivo de publicación en formato impreso y, cada vez más, en formato digital. Se trata de los derechos de reproducir la obra en papel, de distribuir ejemplares en papel y de reproducir y distribuir copias digitales. Ninguno de estos tipos de uso se interpreta naturalmente en el sentido de que incluya el uso de la obra en el entrenamiento de un modelo de IA. Por consiguiente, los proveedores de corpus y los desarrolladores de IA que obtienen licencias de derecho de autor de las editoriales se encuentran en una situación en la que el carácter lícito de las actividades objeto de licencia sigue siendo incierto con respecto a los autores de las obras originales. Incluso cuando el desarrollador ha obtenido una licencia del titular de los derechos en la institución, persiste el riesgo residual de que el autor de la obra original emprenda acciones. Según la jurisdicción, el riesgo no se limita al licenciante: este puede incurrir en responsabilidad por incumplimiento del contrato o de la ley, y el desarrollador, como parte que se beneficia en último término del uso, puede seguir expuesto a reclamaciones por infracción de los autores de las obras originales, a pesar de la licencia. Según la información disponible, en el sector audiovisual los modelos de acuerdos de licencia abordan este riesgo contractualmente: el licenciante declara y garantiza que es titular de todos los derechos necesarios para el uso con fines de entrenamiento. En el caso de los artículos de prensa escritos por periodistas y redactores antes de que, el 1 de junio de 2021, entrara en vigor la modificación pertinente de la Ley de Derecho de Autor de China, el derecho de autor sobre dichos artículos correspondía, por lo general, al periodista o redactor, y no a la agencia de noticias. Esto significaba que ni siquiera una licencia concedida por la agencia de noticias habría amparado el uso del artículo para el entrenamiento de la IA. Este tipo de restricciones puede obstaculizar por sí mismo la aparición de un mercado de licencias voluntarias, ya que una licencia de un intermediario cuyos propios derechos son inciertos añade un nivel más de posible infracción a la cadena de titularidad. También tienen una dimensión distributiva: cuando la concesión de licencias pasa por la intermediación de editoriales, organismos de radiodifusión y repositorios, la asimetría de poder de negociación entre los autores o artistas intérpretes o ejecutantes individuales y dichas instituciones hace que los beneficios de la concesión de licencias puedan no llegar a las personas cuyas obras se utilizan, lo cual es pertinente para elegir entre los modelos de concesión de licencias examinados en la sección B.

Sin embargo, la práctica china pone de manifiesto que, incluso dentro de estas restricciones estructurales, se está produciendo una concesión de licencias voluntarias a gran escala en

[53] Factiva, propiedad de Dow Jones, informó de que, a fines de 2025, aproximadamente 5000 editoriales, de una red de alrededor de 30 000, habían optado por conceder licencias sobre sus contenidos para su uso en sistemas de IA generativa. Véase Dow Jones, “Factiva Sees Growing Publisher Participation in AI Licensing” (Dow Jones, 2025) <https://www.dowjones.com>, consultado el 3 de julio de 2026.

[54] ProRata.ai describe su red, en sus propios términos, como una red integrada por “más de 1000 publicaciones y creadores de contenidos”.

contextos sectoriales y ámbitos específicos. Las emisoras de radio y televisión de Hong Kong se asocian con desarrolladores de IA para proporcionar datos de voz en cantonés de alta calidad. Se suministran novelas digitales a través de la cooperación entre desarrolladores y editoriales en línea que cuentan con un gran número de escritores contratados. Se suministran imágenes a través de la cooperación entre desarrolladores y grandes repositorios de imágenes, que las poseen en virtud de transferencias o licencias exclusivas de los fotógrafos. Se suministran artículos científicos a través de la cooperación entre desarrolladores y operadores de bases de datos académicas, que en algunos casos pasan a ser codesarrolladores del modelo especializado resultante. Algunos proveedores de corpus también producen sus propios materiales de entrenamiento y contratan a fotógrafos, modelos y docentes para que transfieran o concedan licencias sobre las obras resultantes con el fin de usarlas en conjuntos de datos de entrenamiento.

Otras jurisdicciones aplican pautas similares, aunque los mecanismos institucionales son diferentes. En el Brasil, el régimen jurídico de la concesión de licencias voluntarias se rige por las normas generales de la Ley de Derecho de Autor del Brasil. El artículo 29 establece una autorización previa y expresa, y los artículos 4, 31 y 49 imponen una interpretación restrictiva de los negocios jurídicos sobre derecho de autor; por ello, aunque los derechos patrimoniales se hubieran cedido por completo antes de la aparición de la IA generativa, es discutible que esa cesión ampare el uso de la obra para entrenar modelos de IA, puesto que el contrato no podía prever una tecnología que aún no existía. En las jurisdicciones francesa, italiana y otras del sistema de derecho de autor existen doctrinas similares de interpretación estricta.[55]

Cuando la legislación prevé un derecho de autor explícito sobre las puestas en escena, las interpretaciones o ejecuciones u otros objetos de protección, el panorama de la concesión de licencias voluntarias puede ser más sencillo, porque la titularidad de los derechos subyacentes está más clara. Puede haber mayor complejidad en el caso de las categorías de obras creadas por numerosos colaboradores, como las obras audiovisuales y los fonogramas, en las que la incertidumbre de la legislación o de los contratos entre los productores y los colaboradores individuales puede obstaculizar la concesión de licencias voluntarias o desviar sus beneficios de esos colaboradores. La legislación senegalesa, que reconoce expresamente las puestas en escena como obras protegidas,[56] es un ejemplo, aunque la aplicación práctica de ese reconocimiento al entrenamiento de la IA aún no se ha puesto a prueba. En Kenya, la Ley de Derecho de Autor (Cap. 130) reconoce el conjunto habitual de derechos exclusivos y prevé la gestión colectiva del derecho de autor a través de los OGC, pero la infraestructura institucional no se ha concebido para el uso computacional, y la concesión de licencias voluntarias para el entrenamiento de IA aún no es una práctica documentada. En el Canadá, la concesión de licencias voluntarias es el mecanismo por defecto y la negociación contractual es la actividad central. El principal litigio en curso, *Toronto Star Newspapers Ltd v. OpenAI Inc.*, que surgió en el contexto de un presunto uso no autorizado y no de la concesión de licencias propiamente dicha, se examina en la parte III del presente estudio.

Un rasgo habitual de la concesión de licencias voluntarias, individuales o colectivas, es la personalización de las condiciones de la concesión de la licencia al perfil de consumo de datos del sistema para el que se quieren los datos, y dicho perfil varía considerablemente entre las distintas categorías de sistemas de IA generativa que se utilizan actualmente. Cuando los datos se utilizan una sola vez, para el entrenamiento inicial o el perfeccionamiento de un modelo cuyos pesos quedan después fijados, lo habitual es aplicar una tarifa única o una participación en los ingresos por un corpus definido. Cuando el sistema necesita un acceso

[55] Sobre la norma de titularidad anterior a 2021 en la legislación china de derecho de autor respecto de los artículos de prensa escritos por periodistas y redactores, véase la Ley de Derecho de Autor de China (modificada en 2020, en vigor desde el 1 de junio de 2021), artículo 18 (obras por encargo).

[56] Ley de Derecho de Autor del Senegal (Loi n° 2008-09 du 25 janvier 2008), artículo 2. Existen disposiciones comparables en el Acuerdo de Bangui por el que se establece la OAPI (Acta revisada de 2015), Anexo VII.

continuo al material objeto de licencia al generar productos –como la generación aumentada por recuperación (RAG) y los sistemas que realizan un análisis continuo de textos y datos sobre una colección–, el acceso no es puntual, sino continuo. Los modelos de compensación que han surgido en la práctica son el reflejo de esa situación, desde las tasas periódicas hasta el pago por rastreo o inferencia. Aparte del precio, los principales aspectos que deben abordar las partes son el alcance del uso permitido, la duración de la licencia, su ámbito territorial, las consecuencias de su rescisión y la posibilidad de ceder el modelo a terceros. El contexto de la IA plantea la cuestión de si la rescisión implica que el desarrollador debe dejar de utilizar un modelo ya entrenado con el material objeto de licencia, reentrenarlo o destruir las copias almacenadas, opciones que no encajan claramente en las medidas de reparación desarrolladas para las relaciones de licencia anteriores. La rescisión ilustra muy bien la dificultad: en la concesión de licencias tradicional, la rescisión restablece el control del titular de los derechos sobre los usos futuros, pero un titular de derechos que rescinde una concesión de licencias para el entrenamiento de IA no puede, en la práctica, obligar al desarrollador a dejar de utilizar lo que el entrenamiento ya ha fijado en el modelo, en particular cuando el catálogo de que se trata es pequeño en relación con el corpus; la medida de reparación tendría que articularse, en su caso, mediante el reentrenamiento o intentos de desaprendizaje automático, cuyos límites se señalan en la parte I. Incluso cuando dichas intervenciones resultan viables, conllevan costos operativos y de procesamiento que las partes pueden tener que repartir en el propio arreglo, lo que añade una dificultad más a la negociación de este tipo de licencias. Cuando se publican o se venden los pesos de los modelos a usuarios posteriores, se plantea la cuestión de la transmisibilidad: ¿la obligación contractual vinculada a una licencia se traslada a los pesos del modelo? Una cuestión relacionada, pero distinta, es si la remuneración en fases posteriores podría calcularse a partir del propio modelo, por ejemplo, en función de la contribución del material objeto de licencia a los pesos. Estas cuestiones aún no se han resuelto en los contratos que se han hecho públicos y son pertinentes para la posible elección entre los modelos de remuneración examinados más adelante, en la sección B.6.

En el segmento de mayor valor, la concesión de licencias voluntarias en los sectores de la música y el audiovisual se ha materializado principalmente en alianzas individualizadas que combinan la licencia con condiciones comerciales que van mucho más allá del uso de obras protegidas. A lo largo de 2025, las principales compañías discográficas pasaron de los litigios a la concesión de licencias con los principales desarrolladores de herramientas de IA para la creación musical: dos de esos desarrolladores zanjaron las reclamaciones por infracción presentadas contra ellos y acordaron crear nuevas plataformas con licencia así como adquirir un servicio de descubrimiento de conciertos de la compañía discográfica; por otra parte, un servicio de transmisión en continuo acordó con las tres principales compañías discográficas y otros dos titulares de derechos desarrollar productos de IA para la creación musical basados en la adhesión voluntaria.[57] Estos acuerdos extrajudiciales llegaron después de que el sector hiciera públicas varias declaraciones colectivas de principios: en ellas, coaliciones formadas por asociaciones sectoriales, organizaciones de artistas y empresas tecnológicas establecieron que el uso de repertorio protegido para el desarrollo de la IA debe contar con autorización y licencia, y exigieron transparencia en las prácticas de entrenamiento.[58] En el sector audiovisual,

[57] A finales de 2025, tanto Universal Music Group como Warner Music Group llegaron a un acuerdo extrajudicial en sus litigios contra Udio y celebraron acuerdos de licencia para nuevas plataformas; Warner Music Group también llegó a un acuerdo con Suno, que incluyó la adquisición por parte de este último del servicio de descubrimiento de conciertos Songkick. Otro desarrollador, KLAY Vision, firmó acuerdos con las principales compañías discográficas y sus filiales editoriales. En octubre de 2025 Spotify anunció la colaboración en materia de transmisión por flujo continuo con Sony Music Group, Universal Music Group, Warner Music Group, Merlin y Believe.

[58] La Human Artistry Campaign, creada en 2023 por una coalición de más de 150 organizaciones, entre ellas importantes asociaciones de la industria musical y grupos de representación de artistas, establece siete principios fundamentales <https://www.humanartistrycampaign.com>, consultado el 25 de agosto de 2026; los Principios para la creación musical con IA, presentados por Roland Corporation y Universal Music Group y respaldados por más de 50 organizaciones <https://aiformusic.info>, consultado el 25 de agosto de 2026. Ambas iniciativas se describen y

[Sigue la nota en la página siguiente]

en diciembre de 2025 se anunció un acuerdo de licencia por el que un estudio pondría más de 200 de sus personajes a disposición de los usuarios de una plataforma de generación de video y adquiriría una participación de 1000 millones de dólares de los Estados Unidos en el capital del desarrollador; según la información publicada, el acuerdo se abandonó en marzo de 2026, cuando el desarrollador retiró la plataforma.[59] Estos acuerdos atañen en parte a la condición jurídica de los productos, que no forma parte del ámbito del presente estudio; se mencionan aquí únicamente como ejemplos de las estructuras de licencia personalizadas y de alto valor que están surgiendo, y de su fragilidad.

b) Principales modelos de licencias colectivas (voluntarias, ampliadas y obligatorias)

La gestión colectiva voluntaria es la modalidad más común de negociación de licencias colectivas voluntarias. Mediante un contrato, los miembros de un OGC encomiendan a dicho organismo la gestión de determinados derechos; el OGC negocia y suscribe licencias en el marco de su mandato, y los ingresos se reparten entre sus miembros. La mayoría de las jurisdicciones con un sistema de derecho de autor desarrollado cuentan con OGC voluntarios, pero los derechos y las obras que gestionan difieren considerablemente. Las sociedades de gestión de los derechos de ejecución pública de obras musicales son el ejemplo más universal. En China, hay cinco OGC operativos o en proceso de estar plenamente operativos: la Music Copyright Society of China, la China Audio-Video Copyright Association, la China Written Works Copyright Society, la Images Copyright Society of China y la China Film Copyright Association.[60] Los OGC han manifestado públicamente su disposición a conceder licencias para el entrenamiento de la IA, pero, a la fecha del presente estudio, no se ha celebrado en China ningún acuerdo importante de licencia colectiva para el entrenamiento de la IA.[61] La razón estructural, en el contexto chino, es que el mandato del OGC se limita a los derechos de sus miembros, y el desarrollador sigue enfrentándose a un riesgo residual respecto de las obras cuyos autores no son miembros.

En la República de Corea la situación presenta características similares, con la Korea Music Copyright Association y otros organismos parecidos. El carácter limitado del repertorio colectivo parece explicar solo en parte que no se hayan concedido licencias. En virtud de los regímenes normativos actuales, es posible que los desarrolladores tengan pocos incentivos para obtener licencias de forma voluntaria. No obstante, hay otros dos factores que parecen explicar por qué se están celebrando acuerdos en dicho país. En primer lugar, la estructura del mercado permite que un pequeño número de grandes titulares de derechos y proveedores de contenidos negocien directamente con los desarrolladores. En segundo lugar, el riesgo de litigio, con los daños y perjuicios legales previstos que conlleva, constituye un incentivo para llegar a un acuerdo. Los acuerdos de licencia vinculados a acuerdos extrajudiciales descritos en la sección B.2 son ejemplos de ello. Por consiguiente, para explicar por qué la concesión de licencias aún no ha cobrado impulso en China y la República de Corea es necesario sopesar estos factores jurídicos y de mercado junto con las limitaciones del mandato colectivo. A la fecha del presente estudio, en la República de Corea, el Ministerio de Cultura, Deportes y Turismo y la Comisión Coreana de Derecho de Autor estaban trabajando con el objetivo de mejorar el marco normativo.

---

sitúan en un marco más amplio de diseño responsable en de Berardinis *et al.*, “RAIM: Operationalising Trustworthy AI in Music Generation”, citado en la nota a pie 17 de la parte I del presente estudio.

[59] El acuerdo entre The Walt Disney Company y OpenAI, anunciado el 11 de diciembre de 2025, se refería a más de 200 personajes y a una inversión prevista en el capital de 1000 millones de dólares de los Estados Unidos; al parecer, se abandonó después de que, el 24 de marzo de 2026, OpenAI retirara su producto de generación de video Sora, sin que se hubiera llegado a realizar ninguna inversión.

[60] Las tres primeras se suelen abreviar como MCSC, AVCA y CWWCS; la Images Copyright Society of China y la China Film Copyright Association son las dos que aún no están plenamente operativas.

[61] Los principales OGC chinos han manifestado públicamente su disposición a conceder licencias para el entrenamiento de la IA, pero hasta la fecha no se ha celebrado ningún acuerdo de licencia significativo.

Desde el punto de vista institucional, los Estados Unidos de América cuentan con una densa red de organismos sectoriales de gestión colectiva. Las organizaciones de derechos de ejecución (ASCAP, BMI, SESAC, Global Music Rights y AllTrack) gestionan los derechos de ejecución pública de obras musicales; el Mechanical Licensing Collective, creado en virtud de la Ley Orrin G. Hatch–Bob Goodlatte de Modernización de la Música de 2018, gestiona licencias generales de reproducción mecánica para los usos digitales de obras musicales;[62] SoundExchange gestiona las regalías legales por la ejecución digital de grabaciones sonoras en virtud del artículo 114 de la Ley de Derecho de Autor de los Estados Unidos de América; el Copyright Clearance Center, Inc. (CCC) gestiona los derechos de reproducción de materiales textuales publicados; y la Artists Rights Society (ARS) gestiona determinados derechos de reproducción en el ámbito de las artes visuales. Cada organización actúa en su propio sector y en condiciones negociadas con sus representados. En respuesta a la aparición del entrenamiento de la IA como uso de obras protegidas, el CCC ha ampliado su licencia anual para empresas para abarcar determinados usos de la IA y ha anunciado planes para otras licencias específicas para la IA.[63]

Además de China y los Estados Unidos de América, en los dos últimos años los OGC y otros intermediarios han adoptado varias medidas concretas para incluir el entrenamiento de la IA en sus mandatos de concesión de licencias, varias de ellas anunciadas solo a finales de 2025. En Suecia, en septiembre de 2025, la STIM puso en marcha un acuerdo de licencia colectivo en el que participan Songfox, la empresa de música generada por IA, y el proveedor de servicios de atribución Sureel, cuya tecnología permite rastrear los resultados de la IA hasta las obras que influyeron en ellos.[64] En Noruega, Kopinor firmó en diciembre de 2025 un acuerdo, financiado por el Estado noruego, en virtud del cual la Biblioteca Nacional de Noruega puede utilizar contenidos periodísticos para entrenar modelos de lenguaje en noruego.[65] En Dinamarca, el organismo danés de gestión colectiva de publicaciones de prensa firmó, también en diciembre de 2025, una alianza de concesión de licencias basada en la adhesión voluntaria con ProRata AI para el uso de los contenidos de sus miembros en un motor de respuestas basado en la recuperación, mientras que otra organización danesa, Tekst&Node, está desarrollando un modelo de concesión de licencias en el marco del régimen danés de concesión de licencias colectivas ampliadas.[66] En los Países Bajos, la asociación de editoriales privadas de prensa,

---

[62] Orrin G. Hatch–Bob Goodlatte Music Modernization Act, Pub. L. No. 115-264, 132 Stat. 3676 (2018), codificada en 17 U.S.C. artículos 115 y ss. El Mechanical Licensing Collective se designa en virtud del 17 U.S.C. artículo 115.d)3).

[63] Véase www.copyright.com/ai.

[64] STIM, “STIM launches world’s first collective AI music licence” (9 de septiembre de 2025), información publicada por la CISAC, https://www.cisac.org/Newsroom/society-news/stim-launches-worlds-first-collective-ai-music-licence. La licencia basada en la adhesión voluntaria abarca tanto el entrenamiento de modelos como el uso posterior de los resultados de la IA; la compensación combina un pago inicial con una participación en los ingresos del desarrollador, y Sureel, el proveedor de servicios de atribución, rastrea los resultados hasta las obras que influyeron en ellos.

[65] “Kopinor signs historic agreement on newspaper content for AI in Norway” (19 de diciembre de 2025), información publicada por la IFRRO, https://ifrro.org/page/article-detail/ifrro-member-kopinor-signs-historic-agreement-on-newspaper-content-for-ai-in-norway/. El acuerdo, firmado por Kopinor en nombre de la Asociación Noruega de Empresas de Medios de Comunicación, otorga a la Biblioteca Nacional de Noruega acceso a los contenidos periodísticos de más de un año de antigüedad para el entrenamiento de modelos en noruego, y está financiado por el Estado noruego con 45 millones de coronas noruegas anuales. Es distinto del anterior acuerdo “Bookshelf” sobre la difusión digital de libros.

[66] ProRata AI y Danish Press Publications’ Collective Management Organisation (DPCMO), “ProRata Partners with Danish Publishers Group DPCMO” (10 de diciembre de 2025), https://www.prnewswire.com/news-releases/prorata-partners-with-danish-publishers-group-dpcmo-to-launch-the-first-decentralized-sovereign-ai-answer-engine-302637182.html. La DPCMO, que afirma representar al 99 % del sector de la prensa danesa, se adhirió a una alianza de concesión de licencias colectivas basada en la adhesión voluntaria que permite utilizar los contenidos de sus miembros en el motor de respuestas basado en la recuperación de ProRata, según un sistema de reparto de los ingresos. Otra organización danesa (Tekst&Node) está desarrollando un modelo de concesión de licencias de IA en el marco del régimen danés de concesión de licencias colectivas ampliadas, que permite a los titulares de derechos conceder licencias para determinados usos de obras literarias con fines de minería de textos y datos y de IA, con

[Sigue la nota en la página siguiente]

junto con la agencia de noticias ANP, coopera desde julio de 2025 con organizaciones públicas para desarrollar un modelo de lenguaje nacional, GPT-NL.[67] En España, CEDRO se ha sumado a la iniciativa Really Simple Licensing para desarrollar un modelo de concesión de licencias de IA para el mercado español.[68] En el Reino Unido, la Copyright Licensing Agency, junto con Publishers' Licensing Services y la Authors' Licensing and Collecting Society, está desarrollando una licencia colectiva específica para la IA generativa, que abarca el entrenamiento, el ajuste fino y la generación aumentada por recuperación.[69] En otros lugares, las organizaciones de derechos de reproducción han incorporado los derechos de reutilización por sistemas de IA a sus licencias digitales de derecho de autor existentes o están desarrollando soluciones similares: en el Japón, la JAC añadió esos derechos a su Digital Copyright License en abril de 2025;[70] la Copyright Agency de Australia está desarrollando soluciones comparables;[71] y en Alemania, VG Wort ha ampliado su Digital Copyright License a la reutilización por sistemas de IA, mientras que GEMA, el organismo de gestión colectiva, ha propuesto un modelo específico que, a la fecha del presente estudio, aún no había dado lugar a ningún acuerdo.[72] Estos avances siguen siendo desiguales y, en varios casos,

sujeción a la posibilidad de exclusión voluntaria y a la aprobación del Ministerio de Cultura. Véase Terese Foged, "[DK] First extended collective licence approval for text and data mining and the use of works by AI" (IRIS Legal Observations of the European Audiovisual Observatory, 2026) <https://merlin.obs.coe.int/article/10505>, consultado el 3 de julio de 2026.

[67] NDP Nieuwsmedia y TNO, "Large dataset news organizations for Dutch AI language model GPT-NL" (17 de julio de 2025), https://www.ndpnieuwsmedia.nl/2025/07/17/large-dataset-news-organizations-for-dutch-ai-language-model-gpt-nl/. Los miembros de NDP Nieuwsmedia, la asociación de editoriales privadas de prensa de los Países Bajos, junto con la agencia de noticias ANP, están poniendo a disposición partes de sus archivos para entrenar un modelo de lenguaje neerlandés sin ánimo de lucro desarrollado por TNO, NFI y SURF y denominado GPT-NL; las editoriales participantes recibirán una remuneración cuando se publique el modelo.

[68] CEDRO, comunicado de prensa sobre su acuerdo de colaboración con RSL Collective (enero de 2026), https://www.cedro.org/sala-de-prensa/noticias. CEDRO, la entidad española de gestión de derechos reprográficos, se ha incorporado a RSL Collective como socio en España y participa en el comité técnico de RSL, adaptando la norma de concesión de licencias legible por máquina al mercado español para que sus miembros puedan fijar las condiciones de uso de sus obras en el entrenamiento de la IA. Por otra parte, el 16 de diciembre de 2025, CEDRO publicó recomendaciones sobre la concesión de derechos para el entrenamiento de la IA.

[69] Copyright Licensing Agency, "CLA announces development of Generative AI Training Licence" (2025), https://cla.co.uk/development-of-cla-generative-ai-licence/; Publishers' Licensing Services, "PLS and ALCS agree to development of pioneering CLA generative AI licence" (abril de 2025), https://www.pls.org.uk/news-events-policy/news/pls-and-alcs-agree-to-development-of-pioneering-cla-generative-ai-licence/. La CLA, junto con PLS y ALCS, está desarrollando una licencia colectiva de carácter voluntario para el uso de textos en la IA generativa, que abarca el entrenamiento, el ajuste fino y la generación aumentada por recuperación; los titulares de derechos participan registrando sus obras.

[70] Japan Academic Association for Copyright Clearance (JAC) y RightsDirect Japan, "JAC and RightsDirect Japan Announce the Availability of AI Re-Use Rights for Digital Copyright License" (9 de abril de 2025), https://www.businesswire.com/news/home/20250409032666/en/. La JAC añadió derechos de reutilización por sistemas de IA a su Digital Copyright License, una licencia de uso interno para toda la empresa destinada a compañías con sede en el Japón, que se basa en un repertorio de obras de editoriales y sociedades académicas de todo el mundo.

[71] Copyright Agency, "AI and Copyright in Australia" y las preguntas y respuestas que lo acompañan, https://www.copyright.com.au/membership/ai-and-copyright-in-australia/. La Copyright Agency, miembro del Grupo de Referencia sobre Derecho de Autor e IA del Gobierno de Australia, afirma que está desarrollando soluciones de licencia para el uso de material protegido en relación con la IA generativa en Australia.

[72] En el caso de VG Wort, la Digital Copyright License gestionada a través de RightsDirect (filial del CCC) se ha ampliado a la reutilización interna por sistemas de IA; véase https://www.rightsdirect.com/license/vg-wort-digital-license-2/. En 2024, la sociedad de gestión colectiva GEMA propuso un modelo de concesión de licencias para la IA generativa; a la fecha del presente estudio no se había celebrado ningún acuerdo, y paralelamente, GEMA ha emprendido procedimientos por infracción contra desarrolladores de IA (véase *GEMA v. OpenAI*, *supra*). Véase GEMA, "AI and music: GEMA calls for music creators to share in AI providers' revenues and presents first licensing model" (26 de septiembre de 2024) <https://www.cisac.org/Newsroom/society-news/ai-and-music-gema-calls-music-creators-share-ai-providers-revenues-and>, consultado el 3 de julio de 2026; GEMA, "Suno AI and Open AI: GEMA sues for fair compensation" (20 de enero de 2025) <https://www.gema.de/en/news/ai-and-music/ai-lawsuit>, consultado el 3 de julio de 2026.

experimentales, pero indican que en este momento ya no solo se debate la vía de la concesión colectiva de licencias, sino que se está poniendo a prueba activamente.

La concesión de licencias colectivas ampliadas es un modelo de concesión de licencias que surgió en las jurisdicciones nórdicas en la década de 1960 y que otros países también han adoptado. Aborda el problema del riesgo residual extendiendo el efecto de la licencia de un OGC a los titulares de derechos que no son miembros, dentro de un repertorio definido, y prevé un derecho de exclusión voluntaria. Este modelo se ha mencionado en los debates de políticas europeos sobre el entrenamiento de la IA como medio para obtener los derechos sobre grandes corpus textuales y visuales que el desarrollador necesita utilizar, pero respecto de los cuales no resulta práctico conceder licencias individuales. Un estudio del Parlamento Europeo de 2025 sobre la IA generativa y el derecho de autor lo mencionaba expresamente.[73] En los Estados Unidos de América, la Oficina de Derecho de Autor, en la tercera parte de su informe sobre el derecho de autor y la inteligencia artificial, identificó la concesión de licencias colectivas ampliadas como una posible opción intermedia específica para el entrenamiento de la IA, distinta de la licencia obligatoria.[74] La misma propuesta se refleja, con variaciones, en la literatura académica sobre la remuneración: En su obra sobre la IA generativa y la remuneración de los autores, el profesor Senftleben propone un sistema en el que los organismos de gestión colectiva se encargarían de recaudar una remuneración por el uso de servicios de IA generativa cuyas producciones sustituyan a las obras de autores identificables o compitan con ellas, y de repartir lo recaudado entre sus miembros.[75]

Además de los ejemplos ya mencionados, otros dos pueden ilustrar cómo puede adaptarse la gestión colectiva voluntaria al entrenamiento de la IA aprovechando las estructuras institucionales existentes, sin necesidad de esperar a una nueva legislación. En el Brasil, el régimen de ejecución pública de obras musicales está centralizado en una única oficina, el ECAD, que cuenta con respaldo legal y judicial para centralizar la recaudación y distribución de las regalías por ejecución pública de obras musicales y fonogramas, tanto por derecho de autor como por derechos conexos.[76] En el marco de la Ley de Derecho de Autor del Brasil vigente y sin necesidad de modificaciones legislativas, sería jurídicamente posible ejercer de forma colectiva un derecho específico relativo al uso de obras y fonogramas para el entrenamiento de modelos de IA, conforme a un modelo similar al que se aplica actualmente al derecho de ejecución pública de obras musicales. Para ello, los titulares de derechos tendrían que afiliarse a un OGC, y este tendría que prever en sus estatutos la representación del titular de los derechos en relación con otros derechos, y no solo con los derechos de ejecución pública de obras musicales; además, se requeriría la autorización previa del Ministerio de Cultura, el organismo gubernamental encargado de la supervisión de los OGC. El artículo 97 de la Ley de Derecho de Autor del Brasil permite a los OGC ampliar su ámbito de actuación a la concesión de licencias sobre los derechos afectados por el entrenamiento y por la aplicación del modelo (el derecho de reproducción, la comunicación al público o la puesta a disposición y, según la posición adoptada por el Ministerio de Cultura, el derecho de adaptación en la fase de generación de resultados), siempre que se obtengan las autorizaciones necesarias. No obstante, el requisito de recaudación unificada solo se aplica a los derechos de ejecución pública de obras musicales, de modo que, aunque el ECAD o sus afiliados centralizaran en su

[73] Parlamento Europeo, Dirección General de Políticas Interiores, estudio “Generative AI and Copyright: Training, Creation, Regulation”, IUST_STU(2025)774095_EN (2025), en el que se recomienda la concesión de licencias legales como marco óptimo para el entrenamiento de la IA en la UE.

[74] United States Copyright Office, *Copyright and Artificial Intelligence, Part 3: Generative AI Training* (2025).

[75] M.R.F. Senftleben, “Generative AI and Author Remuneration”, 54 IIC 1535 (2023).

[76] Sobre la estructura del sistema brasileño de gestión colectiva y la centralización de la recaudación de las regalías por ejecución pública en el ECAD, véanse la Ley n.º 9610/1998, artículos 97 a 100, y el Decreto n.º 9574/2018. El Supremo Tribunal Federal confirmó la constitucionalidad del modelo en la ADI 2054 (2003) y en la resolución conjunta sobre las ADI 5062 y 5065 (2016).

seno la representación de un derecho relativo al entrenamiento de la IA, ello no impediría la constitución de otros organismos colectivos con el mismo fin.

En Kenya, el Reglamento del Derecho de Autor (Gestión Colectiva) de 2020 regula el registro, el funcionamiento y la supervisión de los OGC.[77] Aunque se trata de un marco de carácter general, que no está específicamente adaptado a los usos computacionales, ofrece una base institucional sobre la que podría construirse la concesión colectiva y voluntaria de licencias para el entrenamiento de la IA. El Anteproyecto de Ley de Derecho de Autor y Derechos Conexos de 2026, publicado para consulta pública por la Junta de Derecho de Autor de Kenya en marzo de 2026, con un plazo de presentación de comentarios que concluyó a finales de ese mes, contiene una disposición (artículo 34) que permite el análisis de textos y datos con fines de investigación no comercial, pero no ofrece una definición legal clara de la expresión "no comercial", no introduce un mecanismo de exclusión voluntaria para los titulares de derechos y no contiene disposiciones específicas ni sobre el uso de obras protegidas por derecho de autor en el entrenamiento de la IA ni sobre la protección por derecho de autor de las obras generadas por IA. La situación de Kenya es ilustrativa de una tendencia más amplia en las jurisdicciones en las que el marco en materia de derecho de autor se está adaptando por etapas: la infraestructura de OGC existente ofrece un posible instrumento para la concesión colectiva de licencias con fines de entrenamiento de la IA, pero la adaptación requiere medidas legislativas o contractuales concretas que aún no se han adoptado.

3. Licencias comunitarias y de custodia para datos lingüísticos primarios

Los mecanismos de concesión de licencias descritos anteriormente se aplican a datos de entrenamiento que ya están fijados y publicados cuando el desarrollador accede a ellos. Existe una categoría distinta de datos de entrenamiento que surge de otra manera. Los modelos de habla y de lenguaje para la mayoría de las lenguas africanas no pueden construirse a partir de corpus textuales extraídos de la web, porque apenas existen textos de ese tipo. En su lugar, los datos se recopilan como datos primarios: grabaciones de audio y transcripciones obtenidas de hablantes mediante trabajo de campo, sesiones de grabación comunitarias y alianzas con organismos de radiodifusión. El derecho de autor nace en el momento de la fijación, y no en el rastreo. La grabación está protegida como grabación sonora, la transcripción como obra literaria, y el hablante puede ser titular de derechos de artista intérprete o ejecutante.[78] Por lo tanto, las cuestiones de autorización y compensación se plantean en el momento de la recopilación, antes de que pueda surgir cualquier cuestión de minería de textos y datos.

Esta categoría de datos también afecta a otro nivel de propiedad intelectual. Las grabaciones de tradición oral, cuentos populares y proverbios son al mismo tiempo obras protegidas por derecho de autor cuya titularidad corresponde a la persona que las fijó y expresiones culturales sobre las que las comunidades son titulares de derechos colectivos en virtud de legislación sui géneris. En Kenya, la Ley sobre la Protección de los Conocimientos Tradicionales y las Expresiones Culturales, n.º 33 de 2016, que da efecto a los artículos 11, 40 y 69.1.c) de la Constitución de Kenya de 2010, reconoce a las comunidades como titulares de derechos sobre sus expresiones culturales, exige el consentimiento fundamentado previo y acuerdos de usuario autorizado para su explotación, y prevé una compensación destinada a las comunidades; el artículo 11.3.a) de la Constitución obliga al Parlamento a garantizar que las comunidades perciban regalías por el uso de sus culturas y de su patrimonio cultural.[79] Cuando

[77] Reglamento de Derecho de Autor (Gestión Colectiva) de 2020 (Kenya). La infraestructura de concesión de licencias se establece en los artículos 33 y 34 y en la parte VII de la Ley de Derecho de Autor (Cap. 130).

[78] Ley de Derecho de Autor (Cap. 130), Leyes de Kenya, artículo 2 (en la que se definen, entre otras cosas, "grabación sonora" y "obra literaria") y disposiciones sobre los derechos de los artistas intérpretes o ejecutantes. Sobre los derechos de los artistas intérpretes o ejecutantes en el plano internacional, véanse los artículos 10 y 14 del WPPT, examinados en la sección B.1, más arriba.

[79] Ley de Protección de los Conocimientos Tradicionales y las Expresiones Culturales, n.º 33 de 2016 (Kenya); Constitución de Kenya de 2010, artículos 11, 40.5 y 69.1.c).

se incorporan datos lingüísticos de este tipo a un corpus de entrenamiento, la cadena de titularidad del derecho de autor y los derechos de la comunidad operan en paralelo, y una licencia de quien fijó la grabación no resuelve por sí misma la situación jurídica de la comunidad prevista en la ley.

En respuesta a esta situación, se ha desarrollado una práctica de concesión de licencias diferenciada. Las licencias abiertas estándar, y en particular las cesiones al dominio público, introducen dichos datos en los procesos de entrenamiento sin proporcionar a la comunidad que los aporta capacidad de gobernanza ni remuneración económica. Las licencias comunitarias cambian esta situación manteniendo la apertura. La licencia de datos abiertos Nwulite Obodo aplica condiciones de uso escalonadas, diferenciadas según el contexto geográfico y de desarrollo del usuario, e incorpora obligaciones de participación en los beneficios; se ha aplicado a conjuntos de datos de voz africanos, como el conjunto de datos de voz DhoNam Dholuo, desarrollado con la Universidad de Maseno.[80] La licencia Esethu, creada para el conjunto de datos de voz en isiXhosa Vuk'uzenzele, permite que las entidades africanas utilicen los datos de forma gratuita, tanto para investigación como con fines comerciales; en cambio, los demás usuarios comerciales deben pagar una licencia, y los ingresos obtenidos se reinvierten en la creación de nuevos datos lingüísticos.[81] La licencia Kaitiakitanga de Te Hiku Media, que regula los datos en lengua maorí, aplica el mismo planteamiento en otra región.[82] Estos instrumentos, junto con la concesión de licencias individuales y colectivas, constituyen un tercer mecanismo mediante el cual se está organizando la autorización y la compensación por el uso de datos lingüísticos primarios en el entrenamiento de la IA, e ilustran cómo puede existir un marco jurídico nacional para los bienes culturales de titularidad colectiva que respalde la práctica contractual. En Kenya, este nivel también interactuará con la excepción relativa al análisis de textos y datos propuesta en el Anteproyecto de Ley de Derecho de Autor y Derechos Conexos de 2026, examinada en el apartado anterior.

4. Excepciones legales relativas a la minería de textos y datos y cláusulas generales

La legislación de derecho de autor de la Unión Europea contiene excepciones explícitas relativas a la minería de textos y datos aplicables a operaciones utilizadas en el entrenamiento de la IA, aunque su aplicación y su idoneidad en ese contexto siguen siendo objeto de controversia. Los artículos 3 y 4 de la Directiva (UE) 2019/790 sobre los derechos de autor y derechos afines en el mercado único digital establecen dos excepciones obligatorias relativas a la minería de textos y datos. El artículo 3 contempla una excepción con fines de investigación para los centros de investigación y las instituciones del patrimonio cultural, sin posibilidad de exclusión. El artículo 4 establece una excepción general de la que puede beneficiarse cualquier persona; esta excepción se aplica siempre que los titulares de derechos no se hayan reservado expresamente el uso de sus obras "de manera adecuada, como medios de lectura mecánica en el caso del contenido puesto a la disposición del público en línea".[83] El Reglamento de IA

[80] Licencia de datos abiertos Nwulite Obodo, versión 1.0 (Data Science Law Lab, Universidad de Pretoria, 2025) <https://datasciencelawlab.africa/nwulite-obodo-open-data-license/>, consultado el 31 de julio de 2026; C.I. Okorie y M. Omino, "Addressing Inequitable Openness in Licences for Sharing African Data and Datasets Through the Nwulite Obodo Open Data Licence" (2025) 7(3) Law, Technology and Humans 94. El conjunto de datos de voz DhoNam Dholuo (Centro de IA Africana de la Universidad de Maseno) se publica con esta licencia: <https://mozilladatacollective.com/datasets/cmjepxo6t08nmmk07iauvua6v>, consultado el 31 de julio de 2026.

[81] Lelapa AI, Way With Words y Data Science for Social Impact (Universidad de Pretoria), "The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages", Actas de la 63.ª Reunión Anual de la Asociación de Lingüística Computacional (2025) https://aclanthology.org/2025.acl-long.1487/, consultado el 31 de julio de 2026.

[82] Te Hiku Media, licencia Kaitiakitanga <https://tehiku.nz/te-hiku-tech/te-hiku-dev-korero/25141/data-sovereignty-and-the-kaitiakitanga-license>, consultado el 31 de julio de 2026.

[83] Directiva (UE) 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, DO L 130/92, artículo 3 (excepción obligatoria con respecto a las extracciones de textos y datos, con fines de investigación científica, realizadas por organismos de investigación e

[Sigue la nota en la página siguiente]

incorpora este marco por remisión obligando a los proveedores de modelos de IA de uso general a establecer directrices para cumplir el derecho de la Unión en materia de derecho de autor y a identificar y respetar las reservas de derechos expresadas de conformidad con el artículo 4.3 de la Directiva sobre los derechos de autor en el mercado único digital, así como a publicar un resumen suficientemente detallado del contenido utilizado para el entrenamiento.[84]

Dos resoluciones alemanas han abordado la aplicación de estas excepciones al entrenamiento de la IA. En *Kneschke v. LAION e.V.*, el Tribunal Regional de Hamburgo resolvió, en septiembre de 2024, que el uso de una fotografía en la construcción del conjunto de datos LAION-5B, que se utiliza en el entrenamiento de modelos de generación de imágenes, estaba amparado por la transposición alemana de la excepción prevista en el artículo 3 de la Directiva sobre los derechos de autor en el mercado único digital relativa a la minería de textos y datos con fines de investigación científica. El tribunal consideró que LAION era un centro de investigación científica en el sentido de la disposición y estimó que el análisis computacional en cuestión constituía minería de textos y datos con fines científicos. También indicó, *obiter dictum*, que la excepción general relativa a la minería de textos y datos del artículo 4 no sería aplicable en ningún caso, porque la reserva de derechos contenida en las condiciones de uso de la plataforma de origen era, a juicio del tribunal, legible por máquina a la luz de las capacidades de los grandes modelos de lenguaje actuales.[85] En apelación, en *Kneschke v. LAION e.V.*, el Tribunal Regional Superior de Hamburgo confirmó la decisión, pero desarrolló con mayor detalle el requisito relativo al formato legible por máquina del artículo 4.3.[86] El tribunal de apelación consideró que el formato legible por máquina exige no solo que la máquina pueda captar la reserva, sino también que un proceso automatizado pueda interpretarla de tal manera que no se procese el contenido al que se refiere. En resumen, la reserva debe ser a la vez legible y procesable por máquina. El tribunal indicó además que el hecho de ofrecer un formato legible por máquina no es una propiedad estática, sino que debe evaluarse en función de las capacidades tecnológicas existentes en el momento del uso controvertido; en este caso concreto, se trataba del estado de la técnica en el segundo semestre de 2021. El razonamiento del tribunal de apelación deja abierta la cuestión de qué se consideraría "legible por máquina" en 2025 o más adelante, cuando existan de forma generalizada modelos capaces de extraer significado de textos en lenguaje natural no estructurado.[87]

instituciones responsables del patrimonio cultural) y artículo 4 (excepción general, con posibilidad de reserva, para la minería de textos y datos de obras a las que se tenga acceso lícito).

[84] Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo, de 13 de junio de 2024, por el que se establecen normas armonizadas en materia de inteligencia artificial (Reglamento de Inteligencia Artificial), DO L, 2024/1689, 12.7.2024, artículo 53.1.c) y d), y considerandos 105 y 106.

[85] *Kneschke v. LAION e.V.*, Tribunal Regional de Hamburgo, 27 de septiembre de 2024 (310 O 227/23) (en la que se resuelve que la construcción del conjunto de datos LAION-5B para su uso en el entrenamiento de modelos de generación de imágenes está amparada por la transposición alemana de la excepción del artículo 3 de la Directiva sobre los derechos de autor en el mercado único digital relativa a la minería de textos y datos con fines de investigación científica (artículo 60d UrhG); el tribunal indicó también, *obiter dictum*, que la excepción general relativa a la minería de textos y datos del artículo 4 no habría sido aplicable, considerando que la reserva de derechos contenida en las condiciones de la plataforma de origen era legible por máquina en el sentido pertinente). Sobre la apelación, véase *Kneschke v. LAION e.V.*, Tribunal Regional Superior de Hamburgo, 10 de diciembre de 2025 (5 U 104/24), examinada más adelante.

[86] *Kneschke v. LAION e.V.*, Tribunal Regional Superior de Hamburgo, 10 de diciembre de 2025 (5 U 104/24).

[87] El Tribunal de Apelación de Hamburgo señaló que el formato legible por máquina exige que la reserva "pueda ser interpretada por una máquina de tal manera […] que, en un proceso automatizado, no se procese el contenido al que se refiere la reserva [de derechos]" (traducción de los autores). Véase también P. Keller, "LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs", Kluwer Copyright Blog (17 y 18 de diciembre de 2025), parte 1.

Aproximadamente un mes antes, en *GEMA v. OpenAI*, el Tribunal Regional de Múnich I había llegado a una conclusión distinta sobre una cuestión relacionada, aunque diferente.[88] GEMA, la sociedad alemana de gestión colectiva, había demandado a OpenAI alegando que ChatGPT había memorizado las letras de nueve conocidas canciones alemanas y que se le podía inducir a reproducirlas. El tribunal de Múnich consideró que tanto la ingestión de las letras en el corpus de entrenamiento como su posterior memorización en el modelo constituían actos de reproducción en el sentido de la transposición alemana de la Directiva InfoSoc; rechazó la aplicación de la excepción general relativa a la minería de textos y datos del artículo 44b de la Ley de Derecho de Autor alemana, por entender que los modelos generativos memorizan las letras de forma permanente, lo que va más allá del análisis de información textual y basada en patrones que la excepción tenía por objeto permitir; y consideró que los resultados generados demostraban que las letras estaban fijadas en el modelo de una manera comparable a una biblioteca digital. El tribunal dictó medidas de cesación, ordenó que se facilitara información y otorgó una indemnización por daños y perjuicios.[89] Esta resolución ha sido recurrida. En algunos círculos esta sentencia se ha interpretado como la primera gran sentencia europea sobre el entrenamiento de la IA favorable a los titulares de derechos; otros la han criticado por reducir a un único acto de reproducción lo que, desde el punto de vista tecnológico, es una secuencia de operaciones distintas, y por aplicar incorrectamente la excepción relativa a la minería de textos y datos al confundir las pruebas de memorización en la fase de generación de resultados con la realidad técnica del proceso de entrenamiento.[90]

Más allá de la jurisprudencia, existe un debate académico sobre si la excepción del artículo 4 de la Directiva sobre los derechos de autor en el mercado único digital es siquiera aplicable al entrenamiento de la IA. Una opinión muy extendida en la doctrina es que el entrenamiento de la IA es una forma de minería de datos amparada por los artículos 3 y 4 de la Directiva sobre los derechos de autor en el mercado único digital. Esta opinión suele invocar el hecho de que el Reglamento de IA remite a dichos artículos. Varios académicos de la UE, como el profesor Tim Dornis, entre otros, han formulado de manera destacada una opinión contraria; en un artículo publicado en la *European Intellectual Property Review* y en un dictamen jurídico más extenso redactado junto con el informático Stober, Dornis sostiene que un examen detenido de los fundamentos tecnológicos del entrenamiento de la IA generativa pone de manifiesto que las operaciones realizadas van más allá de lo que la excepción relativa a la minería de textos y datos tenía por objeto abarcar. En líneas generales, el argumento es el siguiente: la excepción relativa a la minería de textos y datos permite extraer información semántica, estadística y basada en hechos de las obras; el entrenamiento de un modelo de IA generativa, en cambio, extrae y codifica también rasgos sintácticos y expresivos, de una manera que posteriormente se utiliza para generar productos que sustituyen a las obras originales; y la consiguiente interferencia con los intereses legítimos de los titulares de derechos excede el ámbito de lo que la excepción estaba concebida para permitir.[91]

---

[88] *GEMA v. OpenAI*, Tribunal Regional de Múnich I, 11 de noviembre de 2025 (42 O 14139/24), sentencia definitiva. El demandante era GEMA, la sociedad alemana de gestión colectiva; los demandados eran OpenAI OpCo LLC y OpenAI Ireland Limited. El tribunal dictó medidas de cesación, ordenó que se facilitara información y otorgó una indemnización por daños y perjuicios.

[89] Ibid. El tribunal consideró que tanto la ingestión de las letras en el corpus de entrenamiento como su posterior memorización en el modelo constituían actos de reproducción en el sentido del artículo 16 UrhG, que transpone el artículo 2 de la Directiva 2001/29/CE.

[90] Para un comentario crítico sobre este punto, véase G. Frosio, "Copyright in Formaldehyde: How GEMA v OpenAI Freezes Doctrine and Chills AI", Kluwer Copyright Blog (10 y 11 de diciembre de 2025), Partes 1 y 2, donde se sostiene que el tribunal de Múnich confunde las pruebas de memorización en la fase de generación de resultados con la realidad técnica del entrenamiento, y que lo más acertado es considerar que las copias técnicas intermedias lícitas están comprendidas en el ámbito del artículo 5, apartado 1, de la Directiva InfoSoc o en las excepciones relativas a la minería de textos y datos.

[91] T.W. Dornis, "The Training of Generative AI Is Not Text and Data Mining", 47 *European Intellectual Property Review* 65 (2025). Véase también T.W. Dornis y S. Stober, *Urheberrecht und Training generativer KI-Modelle - technologische und juristische Grundlagen, Initiative Urheberrecht* (Nomos, septiembre de 2024).

La profesora Eleonora Rosati esgrime un argumento similar, pero desde un punto de partida diferente. Sostiene que la minería de textos y datos puede formar parte de los procesos de entrenamiento de la IA, pero no es sinónimo de entrenamiento de la IA ni es todo lo que este conlleva; que limitar el debate europeo al artículo 4 de la Directiva sobre los derechos de autor en el mercado único digital es demasiado restrictivo, dada la multiplicidad de actos restringidos por el derecho de autor que el entrenamiento de la IA puede implicar; y que, en cualquier caso, la regla de los tres pasos del artículo 7, apartado 2, de dicha Directiva y del artículo 5, apartado 5, de la Directiva InfoSoc puede excluir el entrenamiento de la IA del ámbito de la excepción, por entrar ese uso en conflicto con la explotación normal de las obras y causar un perjuicio injustificado a los intereses legítimos de los titulares de los derechos. Si esto es así, la concesión de licencias sigue siendo una parte necesaria del análisis.[92] Por su parte, el profesor Matthias Leistner ha señalado que el mecanismo de exclusión voluntaria del artículo 4, apartado 3, probablemente se convertirá en el principal reto de la legislación de derecho de autor de la UE y que, en la actualidad, resulta inviable en la práctica, una posición que se ha recogido en el debate europeo de políticas.[93]

El profesor Senftleben sostiene, por el contrario, que el artículo 4 no solo es aplicable en el contexto del entrenamiento de la IA, sino que también puede ser directamente pertinente a la hora de determinar qué constituye una explotación normal de la obra en el contexto de la IA, lo que sería relevante para la aplicación de la regla de los tres pasos.[94]

La Comisión Europea, en su interpretación del Reglamento de IA, ha considerado aplicable al entrenamiento de la IA el artículo 4 de la Directiva sobre los derechos de autor en el mercado único digital, en el sentido de que se centra en la aplicación del mecanismo de exclusión voluntaria, y no en la exclusión del entrenamiento de la IA del alcance de la excepción. La infraestructura de exclusión voluntaria se aborda en la parte III del presente estudio, en el contexto de la transparencia y la observancia.[95]

El poder legislativo coreano ha considerado una excepción relativa a la minería de textos y datos, pero no la ha aprobado; desde 2021 se han presentado a la Asamblea Nacional proyectos de ley en ese sentido.[96] A falta de una disposición legal, la cuestión del entrenamiento debe evaluarse con arreglo a la cláusula general de uso leal del artículo 35-5 de la Ley de Derecho de Autor de Corea, disposición inspirada en el artículo 107 de la Ley de Derecho de Autor de los Estados Unidos de América e introducida a raíz del tratado de libre comercio entre la República de Corea y los Estados Unidos de América. El debate nacional sigue de cerca las consideraciones examinadas en la presente sección. Los partidarios de una excepción legal destacan la seguridad jurídica que aportaría un texto jurídico expreso; quienes se oponen a dicha excepción responden que el texto jurídico genera sus propias

[92] E. Rosati, “Is text and data mining synonymous with AI training?”, 19 Journal of Intellectual Property Law and Practice 851 (2024); E. Rosati, “The German LAION decision: A problematic understanding of the scope of the TDM copyright exceptions and the transition from TDM to AI training”, IPKat (7 de octubre de 2024); E. Rosati, Copyright Exceptions and Fair Use Defences for AI Training Done for ‘Research’ and ‘Learning’, or the Inescapable Licensing Horizon, SSRN Working Paper No. 5331405 (2025).

[93] M. Leistner, “TDM und KI-Training in der Europäischen Union”, GRUR 2024, 1665 (donde se sostiene que el mecanismo de exclusión voluntaria probablemente se convertirá en el principal reto de la legislación de derecho de autor de la UE y que, en la actualidad, resulta inviable en la práctica).

[94] Senftleben, M. “Text and Data Mining, Generative AI, and the Copyright Three-Step Test”. *IIC 57*, 67, en 79 (2026). https://doi.org/10.1007/s40319-026-01680-2

[95] Reglamento de IA, artículo 53.1.c) (establecerán directrices para cumplir el Derecho de la Unión y para detectar y cumplir, por ejemplo, a través de tecnologías punta, una reserva de derechos expresada de conformidad con el artículo 4, apartado 3, de la Directiva sobre los derechos de autor en el mercado único digital).

[96] Comenzó con el anteproyecto de ley de modificación integral de la Ley de Derecho de Autor de Corea presentado a la Asamblea Nacional el 15 de enero de 2021, que contenía una excepción relativa a la minería de textos y datos; el anteproyecto decayó al término de la vigesimoprimera Asamblea Nacional, y siguen examinándose propuestas posteriores.

ambigüedades interpretativas, como demuestra el litigio alemán sobre el formato legible por máquina antes examinado, que los criterios establecidos en 2024 por el Tribunal Supremo de Corea para la aplicación del artículo 35-5 dotan a la cláusula general de un contenido operativo, y que una excepción amplia podría beneficiar de manera desproporcionada a los desarrolladores extranjeros y desviar valor de los titulares de derechos sobre contenidos en coreano.[97]

En lo que respecta al entrenamiento, la cláusula general ha adquirido un contenido operativo mediante orientaciones administrativas. En febrero de 2026, el Ministerio de Cultura, Deportes y Turismo y la Comisión Coreana de Derecho de Autor publicaron conjuntamente la Guía sobre la aplicabilidad de la doctrina del uso leal al entrenamiento de modelos de IA generativa (mencionada en la parte I del presente estudio), que aplica los cuatro factores del artículo 35-5 a las etapas del proceso de entrenamiento descritas en la parte I del presente estudio. La Guía explica que la forma de obtención del material de entrenamiento es pertinente, de modo que realizar raspados no autorizados, sortear medidas tecnológicas de protección, eludir muros de pago y hacer caso omiso de las reservas expresadas mediante el protocolo de exclusión de robots son factores que pesan en contra de que se aprecie un uso leal, y se ofrecen ejemplos ilustrativos de supuestos a uno y otro lado del límite.[98]

5. Los sistemas de remuneración

a) Tipos de sistemas de remuneración

Cuando se permite el uso de obras protegidas en el entrenamiento de la IA, ya sea en virtud de una excepción, de una licencia o de ambas, puede plantearse la cuestión de la remuneración. Existen varios modelos que se basan en la amplia experiencia de la legislación de derecho de autor en la adaptación a nuevas formas de explotación.

El primero es un sistema de tasas de licencia, negociadas de forma individual o colectiva, que el desarrollador paga en el momento de la constitución del corpus o del entrenamiento. La ventaja de este modelo es que el titular de los derechos recibe un pago por el uso de la obra en el entrenamiento, sin condiciones adicionales; el inconveniente, en el caso de las licencias individuales, es que deja fuera a la larga cola de titulares de derechos individuales para quienes la negociación directa no es viable.

El segundo es un sistema de gravámenes, en el que se impone un gravamen a los desarrolladores de IA o a determinadas categorías de usuarios de servicios de IA generativa, y lo recaudado se reparte entre los titulares de derechos a través de un OGC u organismo similar. El estudio del Parlamento Europeo mencionado en la sección B.3 recomienda una forma de concesión de licencias legales para el entrenamiento de sistemas de IA en la Unión Europea, que combinaría las características de un régimen de gravámenes con la gestión colectiva obligatoria. Se han formulado propuestas comparables en jurisdicciones ajenas a la Unión Europea, a menudo por analogía con los gravámenes por copia privada que existen en muchos países. Han surgido dudas sobre la compatibilidad de dichos regímenes con la regla de los tres pasos del artículo 9.2) del Convenio de Berna, el artículo 10 del WCT y el artículo 13 del Acuerdo sobre los ADPIC y, en la medida en que abarquen derechos distintos del de reproducción, con las condiciones específicas de los artículos 11*bis*.2) y 13.1) del Convenio de Berna.

El tercer modelo, el más reciente y el más singular desde el punto de vista conceptual, es la remuneración por los resultados generados. Según este planteamiento, el objeto de la

[97] Tribunal Supremo de Corea, Decisión 2021Da272001 (2024), en la que se establecen los criterios de referencia para la aplicación del artículo 35-5 de la Ley de Derecho de Autor de Corea.

[98] Guía coreana sobre el uso leal, nota 10 *supra*.

remuneración no es el uso con fines de entrenamiento en sí, sino la generación, por parte del modelo desplegado, de productos que sustituyen a las obras de autores identificables o compiten con ellas. El modelo vincula la remuneración directamente a la posible sustitución de la obra del titular de los derechos en el mercado pertinente. Sin embargo, al parecer, su aplicación requiere mecanismos técnicos para identificar los productos sustitutivos y rastrearlos hasta obras identificables en los datos de entrenamiento.

Se plantean dos cuestiones adicionales. Cuando el número de obras identificadas es muy elevado, la remuneración por obra puede ser tan reducida que resulte meramente simbólica. Además, es preciso aclarar la base de cálculo: en la mayoría de las propuestas, la remuneración no se basaría en la demostración de que un producto determinado reproduce o adapta una obra concreta, sino en la información facilitada por el desarrollador sobre la forma en que se utilizaron las obras en el sistema, planteamiento que presupone los mecanismos de transparencia examinados en la sección C de la presente parte.[99] Por otra parte, los autores pueden verse afectados por la sustitución tanto si su obra se utilizó para el entrenamiento como si no. Como se ha señalado, con la tecnología de filtros actual, en la práctica es poco probable que la mayoría de los productos infrinjan una obra concreta protegida por derecho de autor, de modo que el perjuicio considerado es sistémico, y no una infracción típica del derecho de autor. Una reproducción realizada en la fase de entrenamiento, o que subsista en el modelo, no implica que los productos sean infractores: el análisis de la infracción respecto de cualquier producto depende del parecido de dicho producto con una obra protegida.

La magnitud del problema de la sustitución está cada vez más documentada en el sector de la música: en abril de 2026, Deezer informó de que las piezas sonoras generadas íntegramente por IA representaban aproximadamente el 44 % de sus cargas diarias, es decir, unas 75 000 piezas sonoras al día, mientras que Spotify informó de la retirada de más de 75 millones de piezas sonoras basura en los 12 meses anteriores,[100] y un estudio del sector ha previsto que una parte considerable de los ingresos de los creadores musicales podría verse desplazada por contenidos generados por IA de aquí a 2028.[101] En respuesta, el profesor Senftleben ha sostenido que el criterio de la sustitución podría servir de vehículo conceptual para distinguir el uso legítimo de la IA para respaldar la creatividad humana (que no daría lugar a remuneración) del uso de la IA para sustituir dicha creatividad humana (que sí daría lugar a remuneración). Podría afirmarse que un enfoque similar sostiene la versión del Senado del anteproyecto de ley brasileño n.º 2338/2023, que combina un mecanismo de exclusión voluntaria con un derecho de remuneración que el titular de los derechos puede ejercer de forma individual o colectiva.[102]

---

[99] En el ámbito musical, la integridad de los metadatos relativos a las obras y a los titulares de derechos vinculados al material de entrenamiento se ha señalado como requisito previo de diseño tanto para la atribución como para cualquier reparto de los ingresos derivados de los resultados generados: de Berardinis *et al.*, "RAIM: Operationalising Trustworthy AI in Music Generation", citado en la nota a pie 17 de la parte I del presente estudio.

[100] Deezer, "AI-generated tracks now represent 44% of all new uploaded music" (19 de abril de 2026) <https://newsroom-deezer.com/2026/04/ai-generated-tracks-represent-44-of-new-uploaded-music/>, consultado el 3 de julio de 2026; Spotify, "Spotify Strengthens AI Protections for Artists, Songwriters, and Producers" (25 de septiembre de 2025) <https://newsroom.spotify.com/2025-09-25/spotify-strengthens-ai-protections/>, consultado el 3 de julio de 2026.

[101] Confederación Internacional de Sociedades de Autores y Compositores (CISAC), estudio de PMP Strategy (diciembre de 2024), en el que se prevé que en torno al 24 % de los ingresos de los creadores musicales podría estar en riesgo de aquí a 2028 como consecuencia de la IA generativa, con una cifra más elevada en el caso concreto de los ingresos digitales.

[102] Anteproyecto de ley n.º 2338/2023, aprobado por el Senado Federal del Brasil el 10 de diciembre de 2024 y remitido a la Cámara de Diputados. A la fecha del presente estudio, el proyecto se encuentra ante una Comisión Especial de la Cámara de Diputados; la versión que pueda someterse finalmente a aprobación presidencial aún no es definitiva.

b) Modalidades y base de la remuneración

En el Brasil, el efecto de sustitución de las obras humanas por las creaciones sintéticas puede observarse en sectores concretos, como la ambientación sonora, en la que algunos usuarios han optado por reproducir música generada íntegramente por sistemas de IA con el fin de eludir el pago de las tasas por ejecución pública. Las causas pendientes contra el ECAD sobre esta cuestión proporcionarán, a su debido tiempo, orientación judicial sobre si la ejecución pública de música generada por IA en establecimientos comerciales da lugar a alguna obligación de remuneración. El ejemplo ilustra la interacción entre la remuneración por el entrenamiento de la IA y la remuneración vinculada con los productos de la IA.
En los debates en curso han surgido dos cuestiones sobre un posible sistema de remuneración. La primera cuestión consiste en determinar si la remuneración debería vincularse al entrenamiento de los modelos de IA, al efecto de sustitución (es decir, a los resultados generados) o a ambos. El estudio del Parlamento Europeo y varias propuestas académicas tratan de combinar ambos planteamientos; el anteproyecto de ley brasileña n.º 2338/2023, en su forma actual, se interpreta mejor en el sentido de que vincula la compensación a la ingestión para el entrenamiento: el artículo 65 prevé una compensación por el uso de contenidos protegidos en la minería de datos y en el entrenamiento y el desarrollo de sistemas de IA, de modo que el proyecto establece un mecanismo de compensación vinculado al acto de entrenamiento que no se extiende a la explotación de los productos.[103] La segunda cuestión es si el sistema de remuneración presupone un derecho exclusivo reconocido que haya sido objeto de licencia o de renuncia, o si debe concebirse en su lugar como un derecho sui géneris a una remuneración equitativa, quizá inspirado en la remuneración por copia para uso privado que existe en muchas jurisdicciones.[104] Sin embargo, como se ha señalado, la legislación brasileña vigente podría alcanzar la fase de generación de resultados a través del derecho de reproducción, el derecho de adaptación o el derecho de comunicación al público, aunque no se aprobara el anteproyecto de ley n.º 2338/2023, y, sobre esa base, se requeriría la autorización previa y expresa de los titulares de los derechos.[105] La copia para uso privado, por el contrario, es una forma de compensación, y no una licencia, porque se considera que los actos privados de copia no pueden ser objeto de licencia por su propia naturaleza, por diversas razones, entre ellas la privacidad de los usuarios.

En el Brasil, hay otra observación pertinente para el diseño de los sistemas de remuneración: el efecto de sustitución de las obras humanas por las creaciones sintéticas puede observarse en sectores concretos, como la ambientación sonora, en la que algunos usuarios han tratado de reproducir música generada íntegramente por sistemas de IA con el fin de eludir el pago de las tasas por ejecución pública. Las causas pendientes contra el ECAD sobre esta cuestión proporcionarán, a su debido tiempo, orientación judicial sobre si la ejecución pública de música generada por IA en establecimientos comerciales da lugar a alguna obligación de remunerar a los miembros del ECAD. El ejemplo ilustra la interacción entre la remuneración por el

[103] Esta es la interpretación del artículo 65 del anteproyecto de ley n.º 2338/2023, en la versión aprobada por el Senado Federal, que se adopta en el presente estudio. También se ha interpretado que el anteproyecto de ley combina el criterio vinculado a la ingestión con la atención a la sustitución por los resultados generados, sobre todo porque los criterios relativos a la compensación hacen referencia a los efectos en el mercado. Según la interpretación adoptada en el texto, la compensación está vinculada a la minería de datos, el entrenamiento y el desarrollo.

[104] Un análisis exhaustivo de los posibles regímenes sui géneris y de los usos que quedarían amparados excede el alcance del presente estudio.

[105] Esta es la interpretación del artículo 65 del anteproyecto de ley n.º 2338/2023, en la versión aprobada por el Senado Federal, que se adopta en el presente estudio. También se ha interpretado que el anteproyecto de ley combina el criterio vinculado a la ingestión con la atención a la sustitución por los resultados generados, sobre todo porque los criterios relativos a la compensación hacen referencia a los efectos en el mercado. Según la interpretación adoptada en el texto, la compensación está vinculada a la minería de datos, el entrenamiento y el desarrollo.

entrenamiento de la IA y la remuneración asociada a los resultados de la IA; las decisiones que se adopten en un ámbito pueden influir en el diseño del otro.[106]

Por último, en las jurisdicciones en las que el marco normativo aún está en formación, la cuestión de la remuneración sigue abierta. La Guía sobre la aplicabilidad de la doctrina del uso leal al entrenamiento de modelos de IA generativa de 2026, publicada conjuntamente por el Ministerio de Cultura, Deportes y Turismo y la Comisión Coreana de Derecho de Autor y mencionada en la parte I, se centra principalmente en la cuestión de qué operaciones de entrenamiento pueden realizarse sin licencia y deja la cuestión de la remuneración correspondiente para el proceso de reforma más amplio actualmente en curso.[107] No obstante, gran parte del debate nacional parte de la premisa de que el entrenamiento comercial de modelos de IA debe conllevar una compensación para los creadores, de modo que la cuestión que queda por resolver es el mecanismo de remuneración. Kenya, el Brasil y China se encuentran en situaciones, en cierto modo, comparables: han puesto en marcha un proceso legislativo, las opciones de política se mantienen abiertas y aún no han determinado el modelo que surgirá de dicho proceso. La parte III del presente estudio, dedicada a la observancia, se detiene en las consecuencias prácticas de esta situación jurídica cambiante para los titulares de derechos que tratan de hacer valer sus pretensiones.

PARTE III. OBSERVANCIA

Las operaciones técnicas descritas en la parte I suelen afectar a derechos exclusivos reconocidos por las legislaciones nacionales de derecho de autor y por los tratados internacionales en la materia; los acuerdos de licencia examinados en la parte II ofrecen mecanismos para regularizar el uso de dichos derechos cuando no es aplicable ninguna excepción o limitación. La parte III aborda lo que sucede cuando no se ha dado ninguna de las dos circunstancias: cuando un titular de derechos considera que su obra se ha utilizado en el entrenamiento de un modelo de IA generativa sin autorización y sin el amparo de una excepción aplicable, y trata de hacer valer sus derechos ante los tribunales o mediante procedimientos administrativos.

Hasta hace poco, apenas había jurisprudencia que examinar. Ahora la situación es distinta. A 22 de agosto de 2026, en los Estados Unidos de América se habían interpuesto 137 demandas contra desarrolladores de IA relacionadas con el uso de obras protegidas por sistemas de IA generativa, y se habían registrado unas 30 demandas más en otras jurisdicciones.[108] Tribunales de Alemania, el Brasil, el Canadá, China, los Estados Unidos de América y el Reino Unido, entre otros, han dictado resoluciones que inciden en múltiples aspectos del análisis: si el modelo entrenado es o contiene copias de las obras con las que se entrenó; si las excepciones del derecho europeo se extienden al entrenamiento de la IA; si los tribunales nacionales son competentes respecto del entrenamiento realizado en el extranjero; si los productos que sustituyen a las obras del corpus de entrenamiento afectan al derecho de reproducción o al derecho de puesta a disposición del titular; y si los propios productos pueden

[106] Anteproyecto de ley n.º 2338/2023, artículos 62 y ss. (en la versión remitida a la Cámara de Diputados). La excepción relativa a los secretos comerciales matiza la obligación de transparencia y es objeto de intensos debates en el proceso legislativo.

[107] Ministerio de Cultura, Deportes y Turismo y Comisión Coreana de Derecho de Autor, *Guía sobre la aplicabilidad de la doctrina del uso leal al entrenamiento de modelos de IA generativa* (febrero de 2026); véase la nota a pie 10 de la parte I del presente estudio.

[108] Según el registro de litigios que mantiene el profesor Edward Lee en chatgptiseatingtheworld.com, a 22 de agosto de 2026, el mapa de los Estados Unidos de América recogía 156 demandas y se habían registrado unas 30 demandas más en otras jurisdicciones. Esta cifra está en constante evolución y se indica a la fecha del presente estudio. Véase también el mapa de litigios que mantiene la Copyright Alliance.

ser objeto de protección por derecho de autor (cuestión que queda fuera del ámbito del presente estudio y que se menciona aquí únicamente a título de contexto).

La presente parte se estructura en torno a tres cuestiones principales. En la sección A se extraen enseñanzas del conjunto de la jurisprudencia en su estado actual, organizadas por temáticas y no por jurisdicciones. En la sección B se aborda la cuestión estructural de la observancia de los derechos frente a los desarrolladores de IA, incluidos los aspectos transfronterizos que la complican. En la sección C se señalan los principales problemas prácticos a los que se enfrentan los titulares de derechos que tratan de hacer valer sus derechos, como la transparencia, la proliferación de normas técnicas para la exclusión voluntaria y la dificultad operativa de la obtención de pruebas.

SECCIÓN A. ENSEÑANZAS DE LA JURISPRUDENCIA ACTUAL

La jurisprudencia acumulada en los últimos 18 meses es fragmentaria, varía según la jurisdicción y aún no está asentada doctrinalmente en ningún ordenamiento jurídico. Aún no ha abordado todas las cuestiones. No obstante, arroja luz sobre al menos cuatro cuestiones distintas, en cada una de las cuales los tribunales de las distintas jurisdicciones han empezado a converger o, en algunos casos, a divergir. A continuación se abordan dichas cuestiones sucesivamente.

1. La cuestión del “modelo como copia”

La primera cuestión es si, desde el punto de vista del derecho de autor, el propio modelo de IA entrenado contiene copias de las obras con las que se entrenó. Es una cuestión distinta de la anterior, a saber, si se realizan reproducciones durante la constitución del corpus y el entrenamiento. La naturaleza técnica de la cuestión de la memorización se ha examinado en la parte I del presente estudio. Tres resoluciones dictadas en los últimos 18 meses dan respuestas muy diferentes.

En *GEMA v. OpenAI*, examinado en la parte II, sección B.5, el Tribunal Regional de Múnich I resolvió, en noviembre de 2025, que el entrenamiento de un modelo de lenguaje con letras de canciones y la posterior memorización de dichas letras en el modelo constituían actos de reproducción en el sentido de la transposición alemana de la Directiva InfoSoc, y rechazó la aplicación de la excepción del artículo 4 de la Directiva sobre los derechos de autor en el mercado único digital.[109]

Tan solo unos días antes, en un asunto relacionado con la cuestión del “modelo como copia” que afectaba a una tecnología diferente, el Tribunal Superior de Justicia de Inglaterra y Gales llegó a una conclusión distinta. En la causa *Getty Images (US) Inc. v. Stability AI Ltd.*, la jueza Joanna Smith DBE resolvió, el 4 de noviembre de 2025, que los pesos de Stable Diffusion, un modelo de difusión latente utilizado para la generación de imágenes, no son copias infractoras de las fotografías con las que se entrenó el modelo.[110] Al término del juicio, el demandante renunció a sus pretensiones por infracción directa y por vulneración del derecho sobre bases de datos, al aceptar que las pruebas no demostraban que el entrenamiento se hubiera realizado en el Reino Unido; por lo tanto, el litigio se centró en la infracción indirecta, lo que obligaba al tribunal a determinar si el modelo era un artículo que constituía en sí mismo una copia infractora. El tribunal consideró que las pruebas técnicas presentadas en el juicio demostraban que los pesos de los modelos de difusión no almacenan los datos de

[109] *GEMA v. OpenAI*, Tribunal Regional de Múnich I, 11 de noviembre de 2025 (42 O 14139/24), sentencia definitiva. La resolución se examina también en la parte II, sección B.5.

[110] *Getty Images (US) Inc. v. Stability AI Ltd.*, [2025] Tribunal Superior de Inglaterra y Gales 2863 (Chancery Division) (jueza Smith, 4 de noviembre de 2025). La resolución aborda tanto pretensiones de derecho de autor como de marcas; el presente análisis se refiere únicamente al análisis del derecho de autor. En diciembre de 2025 se concedió autorización para recurrir la resolución sobre la infracción indirecta ([2025] EWHC 3343 (Ch)); el recurso estaba pendiente a la fecha del presente estudio.

entrenamiento y que, a falta de pruebas de memorización en el sentido pertinente, el modelo no era una copia infractora.[111] El tribunal añadió, *obiter dictum*, que, de haber llegado a la conclusión contraria, Stability AI habría incurrido en responsabilidad con arreglo al régimen de la infracción secundaria, porque el desarrollador sabía o tenía motivos para creer que el corpus de entrenamiento contenía obras obtenidas de la web mediante raspado sin consentimiento.[112]

Una tercera resolución, la sentencia del Tribunal de Internet de Hangzhou en el denominado caso "Ultraman de Hangzhou", aborda un conjunto de circunstancias fácticas conexo, aunque distinto. En ese caso, el demandado era una plataforma que alojaba modelos de IA de código abierto a los que tenían acceso los usuarios; algunos usuarios habían entrenado esos modelos con el manga japonés Ultraman, protegido por derecho de autor, y habían subido los modelos entrenados para otros usuarios. El titular del derecho de autor demandó a la plataforma, y el tribunal declaró a la plataforma responsable por contribución a la infracción. En sus *obiter dicta*, el tribunal expresó la opinión de que el acto de entrenamiento en sí podría analizarse como uso leal con arreglo al derecho chino, argumentando que en el entrenamiento de un modelo de IA generativa se utilizan obras protegidas por derecho de autor con el fin de aprender y analizar rasgos lingüísticos, características estilísticas y otros elementos, y no reproducir la expresión creativa original. El tribunal no necesitaba pronunciarse sobre dicha cuestión y el razonamiento no fue determinante para el resultado, pero en sus *obiter dicta* consideró que el entrenamiento podía acogerse al uso leal con arreglo al derecho chino.[113]

Las tres resoluciones no son plenamente coherentes desde un punto de vista puramente técnico, ya que se refieren a distintos tipos de modelos (un modelo de lenguaje basado en transformadores en Múnich, un modelo de difusión latente en Londres y una plataforma que alojaba modelos de generación de imágenes en los que los usuarios en Hangzhou introdujeron ajustes finos) y a distintas situaciones procesales. Tanto el tribunal de Múnich como el de Londres tuvieron que lidiar con la bibliografía informática sobre la memorización en los grandes modelos generativos. Esa bibliografía incluye la línea de trabajo de los profesores Cooper, Grimmelmann, Lemley y otros sobre la extracción de contenido memorizado de modelos de pesos abiertos.[114] Esos mismos trabajos han sido interpretados por el tribunal de Múnich en el sentido de que demuestran que el entrenamiento equivale a una reproducción a gran escala, y por el tribunal de Londres como respaldo de la conclusión de que los pesos de los modelos de difusión no constituyen, por regla general, copias infractoras. La comparación pone de manifiesto que los datos empíricos disponibles sobre la memorización no determinan por sí solos por completo el resultado del análisis jurídico. La cuestión de saber si la capacidad demostrada de un modelo para reproducir secuencias de sus datos de entrenamiento debe analizarse como prueba de que el modelo es una copia del corpus de entrenamiento o como un defecto de comportamiento que debe mitigarse mediante filtros de salida y técnicas de desaprendizaje es una cuestión de calificación jurídica sobre la que distintos tribunales han llegado a conclusiones diferentes a partir de expedientes probatorios que, aunque diferían en su modalidad y en las pruebas efectivamente presentadas ante cada tribunal, se basaban en

---

[111] Sobre la arquitectura de los modelos de difusión, véase la parte I, sección B.1, y las referencias allí citadas.

[112] Para un comentario, véase E. Derclaye, "Getty Images v Stability AI: A landmark judgment reinforcing the need for the UK government to amend its copyright laws", Kluwer Copyright Blog (20 de noviembre de 2025), en línea: https://legalblogs.wolterskluwer.com/copyright-blog/getty-images-v-stability-ai-a-landmark-judgment-reinforcing-the-need-for-the-uk-government-to-amend-its-copyright-laws/.

[113] Tribunal de Internet de Hangzhou, (2024) Zhe 0192 Min Chu No. 1587 (primera instancia); Tribunal Popular Intermedio de Hangzhou (provincia de Zhejiang), (2024) Zhe 01 Min Zhong No. 10332 (segunda instancia).

[114] A. Feder Cooper y J. Grimmelmann, "The Files are in the Computer: Copyright, Memorization, and Generative AI", SSRN Working Paper No. 4803118 (2024). Sobre la investigación empírica relativa a la memorización en los modelos de lenguaje, véase también el estudio de Cooper et al. citado en la nota a pie 18 de la parte I del presente estudio. Véase también M.A. Lemley y A.F. Cooper, "Probabilistic 'Copies' in Generative AI Models", citado en la nota a pie 25 de la parte I del presente estudio.

una bibliografía científica subyacente similar.[115] Tal y como se ha señalado en la parte I, la cuestión de si un modelo contiene una copia de obras protegidas u objetos de derechos conexos debe resolverse modelo por modelo.

La cuestión del "modelo como copia" se ha planteado de otra forma, y con un enfoque procesal distinto, en los Estados Unidos de América, donde algunos demandantes han alegado no solo que el proceso de entrenamiento generó copias, sino que el propio modelo entrenado constituye una obra derivada infractora. En *Kadrey v. Meta Platforms, Inc.*, los autores sostuvieron que los modelos LLaMA no podían funcionar sin la información expresiva extraída de sus libros, que dicha información se conserva en los pesos del modelo y que, por lo tanto, los modelos son obras derivadas no autorizadas, calificación que vincularía la responsabilidad al propio modelo sin necesidad de demostrar una correspondencia entre un producto concreto y las obras con las que se entrenó el modelo. El tribunal rechazó esa teoría en la fase de alegaciones. El demandado acabó imponiéndose por motivos de uso leal, que se examinan en el apartado 4, más adelante, tras alegar que los pesos son parámetros estadísticos derivados de los datos de entrenamiento y no expresión almacenada.[116]

En *Andersen v. Stability AI Ltd.* se planteó una teoría comparable respecto de un modelo de difusión latente: los demandantes describieron el modelo como un medio alternativo de almacenamiento de las copias de las imágenes de entrenamiento; esa calificación se ha enfrentado a dificultades probatorias a falta de un producto que se corresponda estrechamente con una obra protegida, dificultad que se asemeja a la conclusión del tribunal inglés en el caso Getty, según la cual los pesos de un modelo de difusión no almacenan los datos de entrenamiento. La experiencia de los Estados Unidos de América indica que la prueba de que se introdujeron obras protegidas durante el entrenamiento no basta por sí sola para demostrar que el modelo entrenado es una obra derivada de dichas obras.[117] En el momento de redactar el presente estudio, no se había dictado ninguna resolución de apelación sobre ninguna de las cuestiones anteriores.

2. La excepción relativa a la minería de textos y datos

La segunda cuestión, que se plantea en la Unión Europea y en las jurisdicciones cuya legislación de derecho de autor ha recibido la influencia del marco europeo (por ejemplo, los países del Espacio Económico Europeo), es si las excepciones relativas a la minería de textos y datos de los artículos 3 y 4 de la Directiva sobre los derechos de autor en el mercado único digital se extienden al entrenamiento de modelos de IA generativa. El debate doctrinal de fondo se aborda en la parte II, sección B.5, del presente estudio. La jurisprudencia alemana sobre la cuestión se resume aquí con el fin de ser más exhaustivo, prestando especial atención a lo que enseña sobre la observancia de las reivindicaciones de los titulares de derechos.

[115] Para un comentario crítico sobre el razonamiento del tribunal en el asunto GEMA en este punto, véase G. Frosio, "Copyright in Formaldehyde: How GEMA v. OpenAI Freezes Doctrine and Chills AI", Kluwer Copyright Blog (10 y 11 de diciembre de 2025), Partes 1 y 2.

[116] Para conocer la posición de los desarrolladores más allá del contexto del litigio, véase Google, "A Policy Agenda for Responsible Progress in Artificial Intelligence" (libro blanco sobre la regulación de la IA), disponible en https://blog.google/company-news/outreach-and-initiatives/public-policy/white-paper-ai-regulation/, consultado el 7 de julio de 2026.

[117] Sobre la teoría de la obra derivada en los Estados Unidos de América, véanse *Kadrey v. Meta Platforms, Inc.*, n.º 3:23-cv-03417-VC (Distrito Norte de California, juez Chhabria, 20 de noviembre de 2023) (en la que se desestima en la fase de alegaciones la pretensión de que los modelos LLaMA son en sí mismos obras derivadas infractoras, pero se permite que prosiga la demanda por infracción directa basada en la copia con fines de entrenamiento); y *Andersen v. Stability AI Ltd.*, n.º 3:23-cv-00201-WHO (Distrito Norte de California, juez Orrick), resoluciones de 30 de octubre de 2023 y 12 de agosto de 2024 (en las que se examina, entre otras cuestiones, la pretensión de que el modelo almacena copias comprimidas de las imágenes de entrenamiento). La resolución sobre el uso leal en el asunto Kadrey se examina en la sección A.4, más adelante.

Las dos resoluciones alemanas dictadas en la causa *Kneschke v. LAION e.V.* sobre la aplicación de la excepción relativa a la minería de textos y datos al entrenamiento de la IA (en primera instancia, Tribunal Regional de Hamburgo, 27 de septiembre de 2024, y en apelación, Tribunal Regional Superior de Hamburgo, 10 de diciembre de 2025) se examinan en la parte II, sección B.5, más arriba.

De ellas se desprenden dos enseñanzas para la observancia. La primera es que, en las jurisdicciones en las que la excepción relativa a la minería de textos y datos se aplica al entrenamiento de la IA, la posición del titular de los derechos en materia de observancia está condicionada por el mecanismo de exclusión voluntaria. Cuando se ha expresado una exclusión voluntaria válida de una manera que un proceso automatizado pueda detectar y aplicar, el titular de los derechos puede hacerlos valer mediante los procedimientos ordinarios por infracción. Cuando no se ha expresado tal exclusión voluntaria, o cuando la reserva no cumple los requisitos aplicables, puede aplicarse la excepción relativa a la minería de textos y datos, con sujeción a sus demás condiciones, lo que limita las demandas por infracción basadas en el derecho de reproducción. En ese sentido, la capacidad del titular de los derechos para reservárselos depende en parte de la forma técnica de la señal de exclusión voluntaria.

La segunda es que el criterio del formato legible por máquina no es estático. El Tribunal de Apelación de Hamburgo declaró expresamente que el formato legible por máquina debe evaluarse en función de las capacidades tecnológicas existentes en el momento del uso controvertido. En ese caso concreto, se trataba del estado de la técnica en el segundo semestre de 2021, antes de que se generalizaran los grandes modelos de lenguaje capaces de extraer significado de textos en lenguaje natural no estructurados. La resolución indica que una exclusión voluntaria que no era legible por máquina en 2021 puede serlo ahora, y que una exclusión voluntaria que no es legible por máquina en 2025 puede llegar a serlo en el futuro. La consecuencia práctica para la observancia es que los titulares de derechos pueden tener que revisar y, en muchos casos, volver a formular sus exclusiones voluntarias a medida que evoluciona la tecnología; la consecuencia práctica para los desarrolladores de IA es que la cuestión de si la operación era lícita en el momento en que se realizó no queda zanjada por un cambio posterior de las capacidades pertinentes.[118]

3. Competencia respecto del entrenamiento extraterritorial

Una tercera cuestión, de creciente importancia práctica, se refiere a la competencia de los tribunales nacionales respecto de las operaciones de entrenamiento realizadas en infraestructuras situadas en otro país. La principal resolución canadiense sobre la cuestión, dictada en la causa *Toronto Star Newspapers Ltd. v. OpenAI Inc.*, fue emitida el 7 de noviembre de 2025 por el juez Kimmel, del Tribunal Superior de Justicia de Ontario.[119] Una coalición de importantes medios de comunicación canadienses había demandado a OpenAI alegando que el demandado había utilizado sus contenidos para entrenar a ChatGPT, en violación de la legislación canadiense de derecho de autor, de las medidas tecnológicas de protección de dichos contenidos y de las condiciones de uso de los sitios web correspondientes. OpenAI solicitó la desestimación por falta de competencia, alegando que los actos de entrenamiento se habían realizado en servidores situados en los Estados Unidos de América y, por lo tanto, quedaban fuera del ámbito territorial de la Ley de Derecho de Autor del Canadá. El tribunal rechazó ese argumento al considerar que existía una conexión real y

[118] P. Keller, “LAION Round 2: Machine-Readable but Still Not Actionable - The Lack of Progress on TDM Opt-Outs”, Kluwer Copyright Blog (17 y 18 de diciembre de 2025), Partes 1 y 2.

[119] *Toronto Star Newspapers Ltd. v. OpenAI Inc.*, 2025, Tribunal Superior de Justicia de Ontario 6217 (juez Kimmel, 7 de noviembre de 2025). Los demandantes son el Toronto Star, The Globe and Mail, Postmedia y la Canadian Broadcasting Corporation. La solicitud de desestimación por falta de competencia se basaba en el argumento de que los supuestos actos de entrenamiento se habían realizado en servidores situados en los Estados Unidos de América y, por lo tanto, quedaban fuera del ámbito territorial de la Ley de Derecho de Autor del Canadá.

sustancial entre el procedimiento y Ontario y que la cuestión de si los actos alegados afectaban a la legislación canadiense de derecho de autor correspondía al fondo del asunto, y no al análisis de la competencia. El procedimiento pasó a la fase de exhibición de pruebas a principios de 2026; los demandados han recurrido la resolución sobre la competencia, y el recurso estaba pendiente a la fecha del presente estudio.

Aún no se ha determinado el alcance del criterio de la "conexión real y sustancial" en lo que respecta al entrenamiento de la IA. En los comentarios de carácter académico sobre la resolución de la causa Toronto Star se ha señalado que el criterio se desarrolló en el derecho canadiense en el contexto del derecho de comunicación al público, en causas como *SOCAN v. CAIP* y *SOCAN v. ESA*, y que su traslado al derecho de reproducción no es necesariamente sencillo.[120] Tradicionalmente, a efectos de competencia, el derecho de reproducción se ha localizado en el lugar en que se produce la copia o fijación. En la sección B.2 que figura a continuación se aborda la relación entre el alcance territorial del derecho de reproducción y el principio del trato nacional previsto en el artículo 5 del Convenio de Berna.

Otras novedades canadienses ilustran la variedad de respuestas procesales disponibles. En la causa *CanLII v. Caseway AI*, CanLII había demandado en 2024 a Caseway AI, una empresa emergente con sede en Vancouver, por la descarga masiva de su base de datos de resoluciones judiciales con el fin de entrenar un asistente de información jurídica basado en IA. CanLII alegó que, si bien las resoluciones subyacentes son de dominio público, la estructura organizada, los metadatos y la selección de su base de datos constituyen una recopilación protegida por el derecho de autor. En marzo de 2026, las partes llegaron a un acuerdo extrajudicial confidencial; al haber alcanzado un acuerdo, el caso no dio lugar a ninguna resolución judicial sobre el fondo de esas cuestiones. También hay demandas colectivas en curso: en la Columbia Británica, el autor J.B. MacKinnon ha interpuesto demandas colectivas contra Meta, Nvidia, Anthropic y otras empresas, alegando que estas han utilizado copias ilícitas de obras (procedentes del conjunto de datos Books3) para entrenar sus modelos; en Quebec, Anne Robillard ha solicitado la autorización de demandas colectivas contra Meta y OpenAI, en las que reclama indemnizaciones por daños y perjuicios de hasta 20 000 dólares canadienses por obra, el máximo legal para las infracciones con fines comerciales. La misma restricción estructural resultó decisiva en Inglaterra, en el litigio *Getty* examinado en la sección A.1, más arriba, en el que se renunció a las pretensiones por infracción directa cuando las pruebas no demostraron que el entrenamiento se hubiera realizado en el Reino Unido.[121]

En la India, el Tribunal Superior de Delhi confirmó su competencia sobre el entrenamiento realizado en servidores ubicados en Estados Unidos en la causa *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, sosteniendo, a primera vista, que el almacenamiento de las obras en el extranjero no era más que el "último paso de una cadena de acontecimientos" que comienza con el acceso desde la India, y que la objeción del desarrollador era "insostenible, ya que daría lugar a la elusión de la legislación de derecho de autor de la India";[122] el fondo de la sentencia se

---

[120] *Society of Composers, Authors and Music Publishers of Canada v. Canadian Association of Internet Providers*, [2004] Tribunal Supremo del Canadá 45 ("*SOCAN v. CAIP*"); véase también *Society of Composers, Authors and Music Publishers of Canada v. Entertainment Software Association*, [2022] Tribunal Supremo del Canadá 30. En los comentarios de carácter académico sobre la sentencia del caso Toronto Star se ha planteado la observación doctrinal de que el criterio se desarrolló en relación con el derecho de comunicación al público y que quizá no sea fácilmente trasladable al derecho de reproducción.

[121] *CanLII v. Caseway AI* concluyó en marzo de 2026 mediante un acuerdo extrajudicial confidencial. Las demandas colectivas interpuestas por J.B. MacKinnon en la Columbia Británica contra Meta, Nvidia, Anthropic y Databricks, y por Anne Robillard en Quebec contra Meta y OpenAI, se encuentran en la fase de certificación a la fecha del presente estudio. Sobre el máximo legal para las infracciones con fines comerciales, véase la Ley de Derecho de Autor (Copyright Act), RSC 1985, c C-42, artículo 38.1.1.a).

[122] *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, Tribunal Superior de Delhi, procedimiento mercantil (COMM) 1028/2024 (juez Bansal, sentencia de 24 de julio de 2026), 2026 *SCC OnLine Del* 5291, párrafo 51, donde figuran

[Sigue la nota en la página siguiente]

analiza en la sección A.4 más adelante. En el Brasil, la cuestión del entrenamiento extraterritorial se planteó en agosto de 2025, en la demanda interpuesta por Folha de São Paulo contra OpenAI en la que el periódico alegó que el demandado ha utilizado su contenido periodístico para el entrenamiento de ChatGPT y para la prestación de servicios a usuarios brasileños sin autorización.[123] El procedimiento se tramitó bajo secreto judicial; las partes llegaron a un acuerdo en 2026 y no se dictó ninguna resolución sobre el fondo. En la Argentina, Editorial Perfil, una importante editorial de prensa, interpuso en agosto de 2026 una demanda contra OpenAI y Microsoft ante el Tribunal Civil y Comercial Federal, alegando el uso no autorizado de sus contenidos periodísticos para el entrenamiento y el funcionamiento de servicios de IA generativa.[124] De manera más general, se ha formulado otro argumento, aún no sometido a los tribunales brasileños en el contexto de la IA, por analogía con la jurisprudencia del Supremo Tribunal Federal relativa a las plataformas digitales extranjeras que operan en el Brasil.[125] Según este argumento, la oferta de servicios de IA en el Brasil somete el entrenamiento subyacente a la legislación brasileña de derecho de autor, porque la consiguiente generación de productos por parte de usuarios brasileños constituye una continuación del acto de entrenamiento; desde este punto de vista, si el entrenamiento se realizó de forma ilícita, la oferta del servicio que depende de él queda sujeta a la legislación brasileña de derecho de autor en virtud del principio del trato nacional previsto en el artículo 5 del Convenio de Berna.[126] El argumento es, en esencia, una variante del principio según el cual una infracción que comienza en el extranjero, pero produce sus efectos en la jurisdicción correspondiente, puede hacerse valer en dicha jurisdicción, aplicado a un contexto tecnológico en el que la fijación pertinente se produce en una fase anterior al efecto que percibe el usuario.

4. La sustitución

Una cuarta cuestión se refiere a la calificación jurídica de los resultados que sustituyen al material protegido con el que se entrenó el modelo o compiten con él. La cuestión es importante porque determina si el perjuicio económico alegado por los titulares de derechos puede enmarcarse en las categorías existentes del derecho de autor, que se vinculan a los actos de reproducción y de comunicación al público, o si, por el contrario, requiere un mecanismo distinto, examinado en la parte II, el de la remuneración por efectos de sustitución que ningún producto concreto permite reclamar judicialmente. Los fundamentos técnicos se exponen en la parte I, sección C.3.

---

los pasajes citados en el texto. La resolución se dictó en relación con la solicitud de aplicación de medidas cautelares presentada por el demandante; el procedimiento sigue en curso y, en la fecha de este estudio, se había interpuesto un recurso de apelación ante una sala colegiada del Tribunal Superior.

[123] *Folha de São Paulo v. OpenAI*, causa n.º 1107237-96.2025.8.26.0100 (Tribunal de Justicia del Estado de São Paulo). La demanda, que se interpuso en agosto de 2025, se tramita bajo secreto judicial; a la fecha del presente estudio no se había dictado ninguna resolución pública.

[124] Editorial Perfil S.A. c. OpenAI, L.L.C., Microsoft Corporation y Microsoft de Argentina S.R.L. (Tribunal Civil y Comercial Federal, Argentina, demanda interpuesta el 24 de agosto de 2026). A la fecha de redacción se disponía de poca información pública sobre la demanda, pero véase Editorial Perfil, “Perfil demanda a OpenAI y Microsoft por el uso de contenidos periodísticos sin autorización” (24 de agosto de 2026) <https://www.perfil.com/noticias/judiciales/perfil-demanda-a-openai-y-microsoft-por-el-uso-de-contenidos-periodisticos-sin-autorizacion.phtml>, consultado el 29 de septiembre de 2026. Véase también EFE, “El diario argentino Perfil demanda a OpenAI y Microsoft por uso indebido de sus contenidos” (25 de agosto de 2026) <https://www.swissinfo.ch/spa/el-diario-argentino-perfil-demanda-a-openai-y-microsoft-por-uso-indebido-de-sus-contenidos/91953813>, consultado el 29 de septiembre de 2026.

[125] STF, RE 1037396 y RE 1057258 (agrupados). El Supremo Tribunal Federal resolvió que las plataformas pueden incurrir en responsabilidad por los contenidos publicados por los usuarios si no actúan tras recibir una notificación, que las plataformas que prestan servicios en el Brasil deben cumplir la legislación brasileña y que dichas empresas deben contar con un representante legal en el país.

[126] Convenio de Berna para la Protección de las Obras Literarias y Artísticas, artículo 5.1) (trato nacional); véase también el artículo 5.2) (independencia de la protección).

En Estados Unidos, varios tribunales ya han abordado este tema. En *The New York Times Company v. Microsoft Corp.*, ahora agrupado en el litigio por infracción del derecho de autor *In re OpenAI*, el demandado solicitó en una fase temprana del procedimiento la desestimación de las demandas por infracción directa del derecho de autor, alegando que el demandante había utilizado lo que el demandado calificó como *prompts* rebuscados y manipulados para provocar la regurgitación de la que se quejaba. En abril de 2025, el Tribunal de Distrito del Distrito Sur de Nueva York se negó a desestimar esas demandas por ese motivo, y el procedimiento pasó a la fase de exhibición de pruebas. La fase de exhibición de pruebas ha estado dominada por una controversia sobre la presentación de los registros de uso de los consumidores. En una serie de resoluciones que culminaron con la confirmación por el juez Stein, el 5 de enero de 2026, de una resolución anterior de la jueza magistrada Wang, el tribunal ordenó la presentación de una muestra aleatoria y anonimizada de 20 millones de registros de interacciones de usuarios con ChatGPT.[127] La resolución desplaza el foco probatorio de la cuestión de si la regurgitación puede provocarse mediante *prompts* formulados de manera artificiosa a la de con qué frecuencia esta se produce en el uso habitual por parte de los consumidores. El tribunal consideró pertinentes para el análisis del uso leal las pruebas generales y no adversariales sobre los resultados obtenidos por los usuarios para determinar, en particular, si los productos del modelo funcionan como sustitutos en el mercado de las obras del demandante.

Después, en dos sentencias por vía sumaria dictadas por el Distrito Norte de California en junio de 2025, con dos días de diferencia, en las causas *Bartz v. Anthropic PBC* y *Kadrey v. Meta Platforms, Inc.*, dos tribunales federales de distrito llegaron al mismo resultado en la cuestión principal, aunque con un razonamiento sustancialmente distinto. En Bartz, el juez Alsup resolvió que el entrenamiento de un gran modelo de lenguaje con libros adquiridos lícitamente era sumamente transformador y constituía uso leal en virtud del artículo 107 de la Ley de Derecho de Autor de los Estados Unidos de América; al mismo tiempo, resolvió que la descarga y conservación de copias pirateadas de esos libros procedentes de LibGen y PiLiMi no constituía uso leal, y el 17 de julio de 2025 se certificó la demanda colectiva únicamente respecto de las pretensiones relativas a la piratería. Posteriormente, las partes llegaron a un acuerdo extrajudicial por valor de 1500 millones de dólares de los Estados Unidos, que recibió la aprobación definitiva en julio de 2026.128 En *Kadrey*, el juez Chhabria también dictó sentencia por vía sumaria a favor del demandado en la cuestión del uso leal,[129] pero su razonamiento hizo más hincapié en el cuarto factor del uso leal (el perjuicio para el mercado) y menos en el primero (el uso transformador), haciéndose eco de la metodología del Tribunal Supremo en *Andy Warhol Foundation v. Goldsmith*. En comentarios posteriores se ha interpretado que el asunto dependió principalmente de que el demandante no aportó pruebas de sustitución en el mercado, lo que implica que un expediente probatorio distinto podría dar

[127] *In re OpenAI, Inc. Copyright Infringement Litigation*, n.º 1:25-md-03143-SHS-OTW (S.D.N.Y.). El litigio multidistrito agrupa varias demandas interpuestas por organizaciones de noticias, autores y otros demandantes contra OpenAI. Sobre la orden de exhibición de los 20 millones de registros de ChatGPT, véanse las referencias de la nota a pie 20 de la parte I del presente estudio.

[128] *Bartz v. Anthropic PBC*, n.º 3:24-cv-05417 (Distrito Norte de California, juez Alsup, 23 de junio de 2025) (en la que se dicta parcialmente sentencia por vía sumaria a favor del demandado por motivos de uso leal en relación con el entrenamiento con obras adquiridas lícitamente, y se deniega el juicio sumario a favor del demandado en relación con la descarga y conservación de copias pirateadas de obras procedentes de LibGen y PiLiMi). La demanda colectiva se certificó únicamente respecto de las pretensiones relativas a la piratería. El 25 de septiembre de 2025 se aprobó con carácter preliminar un acuerdo extrajudicial por valor de 1500 millones de dólares de los Estados Unidos; la audiencia final sobre la equidad del acuerdo se celebró el 14 de mayo de 2026 ante la jueza Martínez-Olguín, que dejó el asunto pendiente de resolución. La aprobación definitiva estaba pendiente a la fecha del presente estudio. El acuerdo extrajudicial recibió la aprobación definitiva del Tribunal de Distrito en julio de 2026, y el asunto se ha archivado.

[129] *Kadrey v. Meta Platforms, Inc.*, n.º 3:23-cv-03417-VC (Distrito Norte de California, juez Chhabria, 25 de junio de 2025).

lugar a un resultado diferente.[130] Ninguna de las dos resoluciones establece una proposición general según la cual el entrenamiento de la IA es o no es uso leal.

En *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, el Tribunal de Apelación de los Estados Unidos de América para el Tercer Circuito, al confirmar en recurso interlocutorio una sentencia por vía sumaria parcial, dictó la primera resolución de apelación de los Estados Unidos de América sobre el uso leal y el entrenamiento de la IA. El demandado había utilizado los sumarios del demandante, resúmenes editoriales de cuestiones jurídicas elaborados para la plataforma de investigación Westlaw, para entrenar una herramienta de investigación jurídica basada en IA de carácter no generativo. En cuanto al primer factor, el tribunal consideró que el uso era, a lo sumo, mínimamente transformador, ya que servía esencialmente a la misma finalidad de investigación jurídica que el uso que el demandante hacía de los sumarios; esa conclusión, unida al carácter comercial del uso, hizo que el primer factor pesara en contra del uso leal. El tribunal distinguió los casos de copia intermedia por cuanto en ellos la copia era necesaria para acceder a elementos funcionales no protegidos de programas informáticos, mientras que ninguna necesidad comparable justificaba la copia de los sumarios. El segundo factor favorecía al demandado, pero, como señaló el tribunal, rara vez tiene un peso significativo; el tercero pesaba en contra del uso leal, porque se habían tomado sumarios completos sin una justificación transformadora. En cuanto al cuarto factor, el tribunal apreció un perjuicio en tres dimensiones: para el valor de los sumarios como reclamo para las suscripciones a la plataforma del demandante; para el mercado original de las plataformas de investigación jurídica, en el que el demandado competía directamente; y, al término de su análisis del uso leal, para un posible mercado derivado de concesión de licencias sobre sumarios como datos de entrenamiento de IA, que el tribunal calificó de mercado en rápido desarrollo, observando que el demandante utilizaba sus sumarios para entrenar sus propios productos de IA, que su decisión de no conceder licencias a terceros no desmentía la existencia de ese mercado y que la copia no autorizada usurpaba la oportunidad del demandante de acceder a dicho mercado. El tribunal distinguió los modelos generativos en cuestión en los asuntos *Bartz* e *In re OpenAI*, de modo que la evaluación del carácter transformador puede variar en los casos relativos a modelos generativos. No obstante, la resolución confirma, en el plano de la apelación, que un mercado emergente de concesión de licencias sobre contenidos que los creadores normalmente desarrollarían o autorizarían a otros a desarrollar es una consideración pertinente en el marco del cuarto factor.[131] La oleada de acuerdos extrajudiciales en el sector de la música no ha puesto fin a los litigios en ese ámbito. En enero de 2026, un grupo de editores de música interpuso una nueva demanda contra el desarrollador implicado en el litigio Bartz, en la que alegaba el uso de más de 20 000 obras musicales y reclamaba indemnizaciones legales superiores a 3000 millones de dólares de los Estados Unidos, y sociedades de gestión colectiva de Alemania y Dinamarca han seguido litigando contra otro desarrollador de generación de música.[132]

---

[130] Sobre el contraste entre los razonamientos de ambos jueces, véase (entre otros muchos comentarios) el análisis publicado por Wiggin y Dana el 18 de septiembre de 2025, disponible en wiggin.com.

[131] *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.*, n.º 1:20-cv-613 (Tribunal de Distrito de Delaware, juez Bibas, 11 de febrero de 2025) (en el que se concede una sentencia sumaria parcial a favor del demandante por infracción directa y se desestima la defensa basada en el uso leal). La resolución del tribunal de distrito fue confirmada en recurso interlocutorio: n.º 25-2153 (3d Cir., vista celebrada el 11 de junio de 2026, resuelto el 29 de septiembre de 2026) (dictamen de Montgomery-Reeves J.). En una nota a pie, el tribunal de apelación hizo referencia a la declaración de interés presentada por el Departamento de Justicia de los Estados Unidos de América en el litigio por infracción del derecho de autor *In re OpenAI, Inc.*, n.º 1:25-md-3143 (S.D.N.Y., 1 de septiembre de 2026), en la que se sostenía que el entrenamiento de un gran modelo de lenguaje generativo constituye un uso transformador, y distinguió el contexto generativo del caso que tenía ante sí.

[132] Universal Music Publishing Group, Concord y ABKCO interpusieron una segunda demanda contra Anthropic ante el Tribunal de Distrito de los Estados Unidos de América para el Distrito Norte de California el 28 de enero de 2026, en la que alegaban la infracción de más de 20 000 obras. Esta demanda es distinta de la demanda original

[Sigue la nota en la página siguiente]

Fuera de Estados Unidos, los tribunales también se han pronunciado sobre este tema. El caso *Folha de São Paulo v. OpenAI*, descrito anteriormente, planteó la cuestión hasta que las partes llegaron a un acuerdo. Un ejemplo aún pendiente, a la fecha de redacción de este estudio, es el caso del Tribunal de Internet de Beijing en el que cuatro ilustradores han demandado a Xiaohongshu, alegando que la plataforma utilizó sus obras originales para entrenar sin autorización su modelo de IA de generación de imágenes, Trik, y que dicho modelo puede generar imágenes sustancialmente similares a las obras de los demandantes.[133]

En enero de 2025, en la República de Corea, los tres principales organismos de radiodifusión terrestre, KBS, MBC y SBS, demandaron a Naver ante el Tribunal de Distrito Central de Seúl, alegando que sus contenidos informativos se habían utilizado sin autorización para entrenar la familia de modelos de lenguaje HyperCLOVA de Naver. Se trata del primer procedimiento coreano en el que se plantea directamente la cuestión del entrenamiento y, dado el origen común del artículo 35-5 de la Ley de Derecho de Autor de Corea y del artículo 107 de la Ley de Derecho de Autor de los Estados Unidos de América, la jurisprudencia estadounidense sobre el uso leal antes examinada forma parte de los argumentos presentados ante el tribunal coreano.[134] En el Brasil, un actor de doblaje demandó a un centro comercial, alegando que su voz se había utilizado sin autorización en anuncios publicitarios; el centro comercial replicó que la voz en cuestión había sido generada por IA, y este argumento prosperó en primera instancia. Sin embargo, en 2024, el Tribunal de Apelación de São Paulo anuló la sentencia, al considerar que “el uso de la inteligencia artificial, por sí mismo, no elimina (sino que, por el contrario, agrava) el riesgo de uso indebido de los derechos de terceros”, dado que dichas herramientas se entrenan con contenidos extraídos de la web mediante raspado, sin la autorización de los titulares de los derechos. El tribunal también consideró que era necesario reunir más pruebas, y el asunto se devolvió al tribunal de primera instancia; a la fecha del presente estudio no se había dictado una resolución definitiva sobre el fondo.[135]

En la India, por el contrario, la cuestión del entrenamiento ya ha sido objeto del análisis judicial más exhaustivo realizado hasta la fecha fuera de los Estados Unidos y Alemania. En la causa *ANI Media Pvt. Ltd. v. OpenAI OPCO LLC*, una agencia de noticias alegó que sus obras se habían utilizado sin autorización tanto para el entrenamiento de los modelos de lenguaje de la parte demandada como para generar resultados. Al pronunciarse sobre la solicitud de medida

---

interpuesta por las editoriales contra Anthropic en 2023 en relación con letras de canciones, que se presentó en el Distrito Medio de Tennessee y posteriormente se remitió al Distrito Norte de California. Véanse Music Business Worldwide, “UMG, Concord and ABKCO sue Anthropic for $3bn in what could be single-largest non-class action copyright case” (28 de enero de 2026) <https://www.musicbusinessworldwide.com/umg-concord-and-abkco-sue-anthropic-for-3bn-in-what-could-be-single-largest-non-class-act/>, consultado el 3 de julio de 2026; sobre el caso de 2023, Rolling Stone, “Music Publishers Sue Amazon-backed AI Platform for Copyright Infringement” (18 de octubre de 2023) <https://www.rollingstone.com/music/music-news/umg-concord-abkco-sue-anthropic-ai-company-copyright-infringement-1234856761/>, consultado el 3 de julio de 2026; Music Business Worldwide, “Universal, Concord and ABKCO ask court for injunction to stop Anthropic AI from using their lyrics” (16 de noviembre de 2023) <https://www.musicbusinessworldwide.com/universal-concord-and-abkco-ask-court-for-injunction-to-stop-anthropic-ai-from-using-their-lyrics/>, consultado el 3 de julio de 2026. GEMA (Alemania) y Koda (Dinamarca), entre otras, han seguido litigando contra Suno pese a los acuerdos extrajudiciales alcanzados por las compañías discográficas con ese desarrollador y con Udio a finales de 2025.

[133] Tribunal de Internet de Beijing (demanda presentada en noviembre de 2023; pendiente a la fecha del presente estudio); los cuatro demandantes son ilustradores que alegan que Xiaohongshu utilizó sus obras originales para entrenar sin autorización el modelo de IA de generación de imágenes Trik.

[134] *KBS, MBC y SBS v. Naver Corp.*, Tribunal de Distrito Central de Seúl (demanda presentada el 13 de enero de 2025). Las pretensiones se basan en la Ley de Derecho de Autor de Corea y en la Ley de Prevención de Competencia Desleal, y los demandantes reclaman una indemnización por daños y perjuicios y medidas cautelares para que cese todo uso ulterior. Los mismos organismos de radiodifusión interpusieron una demanda comparable contra OpenAI ante el mismo tribunal en febrero de 2026.

[135] Tribunal de Justiça de São Paulo, recurso de apelación civil n.º 1119021-41.2023.8.26.0100 (6.ª Sala de Derecho Privado, magistrado ponente Costa Netto, sentencia de 31 de octubre de 2024), presentado en el Foro de Jueces de Propiedad Intelectual de la OMPI de 2025; véase <https://www.wipo.int/wipolex/en/judgments/details/3217>, consultado el 19 de septiembre de 2026.

cautelar presentada por la agencia, y tras haber confirmado su competencia territorial sobre el entrenamiento impartido a través de servidores ubicados en los Estados Unidos (véase la sección A.3 anterior), el Tribunal Superior de Delhi dictaminó que el entrenamiento afecta al derecho de reproducción, ya que el artículo 14.a)i) de la Ley de Derecho de Autor de 1957 extiende dicho derecho al almacenamiento de una obra "en cualquier soporte por medios electrónicos" y la eliminación de las copias tras el entrenamiento no anula dicho almacenamiento. No obstante, el tribunal consideró, a primera vista, que el almacenamiento entraba dentro del ámbito de aplicación de la disposición sobre uso leal del artículo 52.1)a), que abarca la utilización con fines de "uso privado o personal, incluida la investigación". El término "investigación" se interpretó desde una perspectiva actualizada: "El proceso de entrenamiento de los grandes modelos de lenguaje que sustentan ChatGPT puede considerarse una forma de investigación destinada a generar nuevos conocimientos" y "los actos de investigación ulteriores no pueden circunscribirse exclusivamente a actuaciones humanas", porque "en última instancia, la investigación se lleva a cabo a instancias de los seres humanos y en beneficio de estos". El carácter comercial del demandado tampoco situaba dicho uso fuera del ámbito de aplicación de la disposición: el tribunal calificó el uso de "puramente privado", ya que las copias de entrenamiento permanecían en un entorno cerrado accesible para las modelos, pero no para el público, y el artículo 52.1)a) no contiene ninguna exclusión de los usos comerciales.

En cuanto al requisito de la equidad, el tribunal consideró que el uso de las obras se limitaba al entrenamiento de los modelos, que en las actuaciones no constaba prueba alguna de sustitución en el mercado ni de pérdida de ingresos más allá de meras alegaciones sin respaldo, y que los beneficios públicos de la tecnología debían ponderarse en el análisis; también observó que el demandante podría haber bloqueado los rastreadores del demandado, pero no lo había hecho. La medida cautelar fue denegada además por considerar que cualquier perjuicio, en caso de que se demostrara en el juicio, sería indemnizable mediante una indemnización por daños y perjuicios; el tribunal señaló que la obligación de obtener licencias de todas las fuentes de noticias comparables haría que el desarrollo de grandes modelos de lenguaje resultara económicamente inviable. La resolución es de carácter provisional: el procedimiento sigue en curso y se dirige a la fase de juicio; además, en la fecha de este estudio estaba pendiente un recurso de apelación ante una sala colegiada. No obstante, su interés comparativo es considerable. Mientras que los tribunales alemanes se han planteado si una excepción específica para la minería de textos y datos con fines concretos se aplica al entrenamiento de la IA, y los tribunales de los Estados Unidos se han preguntado si el entrenamiento constituye un uso leal, el Tribunal Superior de Delhi abordó la misma cuestión a través de una cláusula de "acto leal" que enumera "el uso privado o personal, incluida la investigación", y respondió, en esta fase, a favor del desarrollador de IA, basándose en un razonamiento (la ausencia de un perjuicio demostrado para el mercado, la disponibilidad de una opción de exclusión voluntaria no utilizada, los beneficios públicos de la tecnología), que guarda paralelismo con los factores sopesados en las resoluciones estadounidenses analizadas anteriormente.

SECCIÓN B. OBSERVANCIA FRENTE A LOS DESARROLLADORES DE IA

La jurisprudencia examinada en la sección A ilustra la evolución sustantiva de los últimos 18 meses. En la sección B se aborda la cuestión estructural de cómo pueden los titulares de derechos hacer valer sus pretensiones frente a los desarrolladores de IA, habida cuenta de los aspectos transfronterizos de la tecnología y de las asimetrías entre las partes que suelen intervenir en este tipo de litigios.

1. Rasgos estructurales del panorama de la observancia

Existen tres rasgos estructurales recurrentes. El primero es la persistencia y la agudeza del problema transfronterizo señalado en la sección A.3, más arriba. El entrenamiento de un modelo de IA generativa suele realizarse en infraestructuras situadas en una jurisdicción

distinta de la del titular de los derechos; a continuación, el modelo se despliega y se utiliza simultáneamente en muchas jurisdicciones, y los productos se generan en la jurisdicción de cada usuario. Cada una de esas fases puede quedar sujeta a la legislación de un país distinto. La jurisprudencia aún no ha dado una respuesta uniforme a la cuestión de cómo interactúan esos distintos regímenes jurídicos nacionales, y los desarrolladores pueden enfrentarse a obligaciones incoherentes en las distintas jurisdicciones en las que ofrecen sus servicios.
El segundo es el papel del principio del trato nacional previsto en el artículo 5.1) del Convenio de Berna. En virtud de este principio, los autores gozan, en los países de la Unión distintos del país de origen de la obra, de los derechos que las leyes respectivas conceden a sus nacionales, así como de los derechos especialmente concedidos por el Convenio. Este principio se basa en la no discriminación y, por sí mismo, no determina el ámbito de aplicación territorial de la legislación nacional.

El tercero es el recurso a mecanismos procesales de agrupación de demandas, cuyo principal ejemplo es el procedimiento multidistrito de los Estados Unidos de América. La agrupación de las demandas interpuestas contra OpenAI por organizaciones de noticias, autores y otros demandantes en un único litigio multidistrito, en el Distrito Sur de Nueva York, ha centralizado procedimientos que, de otro modo, se habrían dispersado en acciones individuales. Del mismo modo, la demanda colectiva Bartz ha agrupado demandas relativas a copias pirateadas de aproximadamente 482 000 libros y ha dado lugar a un acuerdo extrajudicial de considerable magnitud, como se señaló anteriormente. En las jurisdicciones en las que no existen mecanismos de demanda colectiva (“acción colectiva”) o de procedimiento multidistrito, o en las que están menos desarrollados, el costo de la observancia individual puede resultar prohibitivo, en particular para los autores individuales y los pequeños titulares de derechos. Las demandas colectivas de Quebec descritas en la sección A.3 constituyen un intento de aplicar el mismo mecanismo en un entorno procesal distinto; las demandas colectivas de la Columbia Británica presentan características comparables. En las jurisdicciones que carecen de estos mecanismos, la capacidad práctica de los titulares de derechos individuales para hacer valer sus derechos frente a desarrolladores con abundantes recursos se ve considerablemente limitada.

2. Asimetrías entre las partes

Además de estos rasgos, la jurisprudencia pone de manifiesto varias asimetrías que complican la observancia en la práctica. La primera es la asimetría de información: los desarrolladores disponen de los registros técnicos (manifiestos de los conjuntos de datos, registros de rastreo, configuraciones de entrenamiento) a partir de los cuales puede, en principio, determinarse la composición del corpus de entrenamiento, aun cuando esos registros sean incompletos; los titulares de derechos no tienen un acceso comparable, salvo cuando la extracción u otras demostraciones empíricas aportan pruebas indirectas. Esta cuestión se aborda más adelante, en el examen de la transparencia de la sección C.1. La segunda es la asimetría de recursos: los desarrolladores, en particular los grandes desarrolladores de modelos de vanguardia, disponen de recursos superiores a los de la mayoría de los titulares de derechos. El costo de emprender una acción para la defensa de los derechos prolongada, con el trabajo pericial técnico que hoy se espera en este tipo de casos, es considerable. La tercera es la incertidumbre doctrinal: cuando el derecho no está asentado, la jurisprudencia publicada puede verse condicionada por efectos de selección, ya que las partes resuelven algunas controversias mediante acuerdo y litigan otras, mientras que la capacidad de los titulares de derechos para llevar a cabo casos de prueba se ve limitada por los costos procesales. Por consiguiente, la jurisprudencia resultante debe valorarse teniendo presentes esos efectos de selección.

SECCIÓN C. DESAFÍOS EN LA OBSERVANCIA DE LOS DERECHOS

En el estado actual del derecho y de la tecnología, los titulares de derechos que tratan de hacer valer sus derechos frente a los desarrolladores de IA se enfrentan a tres problemas principales. Se abordan sucesivamente a continuación.

1. Transparencia

La escasa transparencia sobre el proceso de entrenamiento constituye el primer desafío. Salvo contadas excepciones, los desarrolladores de IA no han divulgado la composición de sus corpus de entrenamiento; los resúmenes facilitados, cuando los hay, son tan generales que no permiten a un titular de derechos determinar si se utilizó su obra concreta. Esta opacidad se ve reforzada por la invocación del secreto comercial: a principios de 2026, por ejemplo, el desarrollador de un sistema de generación de música solicitó que se mantuviera bajo secreto incluso el tamaño de su conjunto de datos de entrenamiento en el litigio promovido por grandes compañías musicales, alegando un perjuicio competitivo.[136] Uno de los motivos es que la composición del corpus de entrenamiento puede constituir un activo comercial, y la divulgación exhaustiva de su contenido puede afectar a la posición competitiva del desarrollador. Como resultado de ello se produce una asimetría de información que puede limitar la capacidad práctica de los titulares de derechos para detectar posibles infracciones y emprender acciones de observancia.

La Unión Europea ha abordado esta cuestión mediante un requisito legal expreso de divulgación. El artículo 53.1.d) del Reglamento de IA obliga a los proveedores de modelos de IA de uso general a elaborar y poner a disposición del público un "resumen suficientemente detallado" del contenido utilizado para el entrenamiento del modelo, con arreglo al modelo facilitado por la Oficina de IA de la UE.[137] Su aplicación depende en parte de qué grado de detalle se considere suficiente a efectos de la disposición. Además, la disposición está matizada por un reconocimiento explícito de la necesidad de respetar los secretos comerciales y la información empresarial confidencial, y la aplicación del requisito es actualmente objeto de trabajos de política y normalización en curso.[138] Esta tensión entre la obligación de divulgación y la excepción relativa a los secretos comerciales es objeto de intensos debates. Es probable que el alcance operativo final de la obligación dependa tanto de la aplicación del modelo publicado por la Oficina de IA en julio de 2025 como de la interpretación que las autoridades judiciales y de otro tipo de cada país den a esa excepción.

Otras jurisdicciones han abordado la transparencia mediante instrumentos diferentes. En China, los contenidos generados por IA están sujetos a un amplio régimen de obligaciones de etiquetado.[139] El régimen chino de etiquetado no está dirigido principalmente a la protección del derecho de autor, sino a la prevención del fraude, el engaño y la confusión del público; no obstante, en la práctica ha permitido a los titulares de derechos cierto grado de inferencia retrospectiva, ya que el etiquetado de los contenidos generados por IA permite identificar determinados productos como generados por IA e investigar, a partir del producto, si es probable que se hayan utilizado obras protegidas en el entrenamiento.

En la República de Corea, la Ley Marco sobre el Desarrollo de la Inteligencia Artificial y el Establecimiento de una Base de Confianza, en vigor desde el 22 de enero de 2026, exige que se informe previamente a los usuarios de los servicios de IA generativa y que se etiqueten los

[136] Music Business Worldwide, "Suno moves to keep size of its AI training data sealed in UMG and Sony's copyright case, citing competitive harm" (2026) <https://www.musicbusinessworldwide.com/suno-moves-to-keep-size-of-its-ai-training-data-sealed-in-umg-and-sonys-copyright-case-citing-competitive-harm/>, consultado el 7 de julio de 2026.

[137] Reglamento (UE) 2024/1689, artículo 53.1.d), mencionado en la nota a pie 82 de la parte II del presente estudio.

[138] Reglamento de IA, considerando 107, en el que se deja claro que la obligación de publicar un resumen debe cumplirse de manera que se respeten los secretos comerciales y la información empresarial confidencial, sin exigir su divulgación.

[139] Disposiciones sobre la Administración de los Servicios de Recomendación Algorítmica de Información en Internet (en vigor desde marzo de 2022); Disposiciones sobre la Gestión de la Síntesis Profunda en los Servicios de Información en Internet (en vigor desde enero de 2023); Medidas Provisionales para la Administración de los Servicios de Inteligencia Artificial Generativa (en vigor desde agosto de 2023); Medidas para el Etiquetado de Contenidos Generados y Sintetizados mediante Inteligencia Artificial; y la norma nacional obligatoria del mismo título aprobada por la Administración Estatal de Regulación del Mercado en 2025.

contenidos producidos por IA.[140] Las obligaciones de etiquetado, al igual que las de China, no tienen por objeto principal proteger el derecho de autor; sin embargo, pueden facilitar el mismo tipo de inferencia retrospectiva a partir de los productos. Las cuestiones de derecho de autor se abordan en la Ley de Derecho de Autor de Corea y no en la legislación marco, una distribución entre instrumentos que refleja la relación existente en la Unión Europea entre el Reglamento de IA y el acervo en materia de derecho de autor.

En el Brasil, la respuesta legislativa figura en el anteproyecto de ley n.º 2338/2023, que, en su forma actual, obliga a los desarrolladores de IA a divulgar el uso de materiales protegidos en el desarrollo de sus sistemas mediante la publicación de un resumen; la obligación está sujeta a la salvedad de que deben respetarse los secretos comerciales e industriales.[141] El efecto práctico de la obligación de divulgación dependerá de la amplitud con que se interprete la excepción relativa a los secretos comerciales. La misma observación es aplicable, *mutatis mutandis*, al marco europeo.

Conviene realizar dos observaciones sobre la transparencia como cuestión de observancia. La primera es que la transparencia puede facilitar la observancia, pero no basta por sí sola. La publicación de un resumen, por detallado que sea, no da lugar por sí misma a una causa de acción; el titular de los derechos debe seguir demostrando que se utilizó su obra, que el uso afectó a sus derechos exclusivos y que no es aplicable ninguna excepción. La obligación de transparencia reduce la asimetría de información, pero no la elimina. La segunda es que el diseño de la obligación de transparencia interactúa con el diseño del mecanismo de exclusión voluntaria. Un titular de derechos que haya expresado su exclusión voluntaria y cuya obra figure en el resumen de transparencia como incluida en el corpus de entrenamiento puede tener una base más clara para una reclamación de observancia, con sujeción a la legislación aplicable y a las demás condiciones de la excepción pertinente; la cuestión práctica es si el resumen será lo bastante específico como para permitir esa identificación.

2. Ataques a los datos, filtrado y proliferación de normas de exclusión voluntaria

El segundo desafío se refiere a las técnicas con las que titulares de derechos, desarrolladores e intermediarios han tratado de abordar las características técnicas del entrenamiento y la inferencia que inciden en la observancia. Pueden distinguirse cuatro categorías de técnicas.

La primera son los ataques de extracción, desarrollados en la bibliografía sobre investigación en seguridad e invocados ahora en los litigios. Los ataques de extracción son técnicas que permiten inducir a un modelo entrenado a reproducir contenidos de sus datos de entrenamiento mediante *prompts* cuidadosamente elaborados. Los trabajos en este ámbito, examinados en la parte I del presente estudio, se han citado en la bibliografía sobre seguridad como prueba de que la extractabilidad puede ser una propiedad estructural de los grandes modelos generativos, y no solo el resultado de una determinada estrategia de formulación de *prompts*.[142] Desde el punto de vista de los titulares de derechos, los ataques de extracción son una técnica de obtención de pruebas: pueden aportar pruebas de que determinadas obras se incluyeron en el entrenamiento o pueden recuperarse de un modelo. Desde el punto de vista de los desarrolladores, esos mismos ataques constituyen una vulnerabilidad que debe mitigarse mediante la deduplicación de los datos de entrenamiento, el filtrado de salida y las técnicas de desaprendizaje. La cuestión doctrinal de si la forma en que se obtiene un producto memorizado debe influir en el análisis sustantivo del derecho de autor se encuentra actualmente ante los

[140] Ley Marco sobre el Desarrollo de la Inteligencia Artificial y el Establecimiento de una Base de Confianza (República de Corea), promulgada en enero de 2025 y en vigor desde el 22 de enero de 2026.

[141] Anteproyecto de ley n.º 2338/2023 (Brasil), artículo 62 (en la versión remitida a la Cámara de Diputados); sobre la situación legislativa del anteproyecto de ley, véase la nota a pie 100 de la parte II del presente estudio.

[142] Sobre los ataques de extracción, véanse la nota a pie 18 de la parte I (Cooper, Gokaslan, Ahmed, Cyphert, De Sa, Lemley, Ho y Liang) y la nota a pie 22 (Carlini et al. (2021) y Nasr et al. (2025)).

tribunales de varias jurisdicciones, como se explica en la parte I, sección C.2, del presente estudio.

La segunda es la familia de los filtros de salida y de los filtros de entrada. Los desarrolladores han implantado con creciente frecuencia filtros que impiden que el modelo reproduzca determinados contenidos identificados (filtros de salida) o que excluyen determinados contenidos del corpus de entrenamiento en la fase de preparación de los datos (filtros de entrada).[143] El papel de esos filtros en el análisis sustantivo no está resuelto. En los Estados Unidos de América, las resoluciones *Bartz* y *Kadrey* han abordado el papel de los filtros de distintas maneras; en particular, el juez Chhabria consideró que la cuestión de la sustitución en el mercado dependía en parte de la eficacia de los filtros para impedir que el modelo generara productos que sustituyeran a obras originales concretas. En la Unión Europea, las obligaciones de transparencia del Reglamento de IA hacen referencia a las políticas que los desarrolladores mantienen en materia de derecho de autor; la aplicación de dichas obligaciones puede incidir en la calificación jurídica de las prácticas de filtrado como parte del cumplimiento general de un desarrollador.

La tercera es la familia de técnicas de envenenamiento de datos, mediante las cuales los titulares de los derechos o terceros introducen datos maliciosos, erróneos o alterados en los conjuntos de datos con el fin de comprometer el rendimiento de cualquier modelo entrenado con ellos. En China, la Administración del Ciberespacio ha tratado el envenenamiento de datos como una amenaza para la seguridad y ha puesto en marcha una campaña especial para combatirlo (Operación Qinglang para rectificar el caos de las aplicaciones de IA); la norma nacional recomendada relativa a los requisitos básicos de seguridad para los servicios de inteligencia artificial generativa exige el filtrado de los datos de entrenamiento para eliminar la información ilegal y nociva.[144] En las demás jurisdicciones examinadas aún no ha surgido un tratamiento normativo comparable del envenenamiento de datos. La importancia práctica del envenenamiento de datos como mecanismo de autotutela para los titulares de derechos todavía no se ha evaluado con rotundidad en ninguna jurisdicción; las cuestiones técnicas y jurídicas que plantea siguen, en esencia, pendientes, a saber: si el uso de técnicas de envenenamiento contra los desarrolladores de IA constituye una infracción reconocible de algún derecho del desarrollador o un ejercicio defendible de los propios derechos del titular.

La cuarta es especialmente relevante para la observancia en Europa. Se trata de la cuestión de las normas relativas a la expresión de las exclusiones voluntarias en formato legible por máquina.[145] La resolución del Tribunal de Apelación de Hamburgo en la causa *Kneschke v. LAION e.V.*, examinada en la parte II, sección B.5, más arriba, establece que una exclusión voluntaria válida con arreglo al artículo 4, apartado 3, de la Directiva sobre los derechos de autor en el mercado único digital debe ser a la vez legible y procesable por máquina. La aplicación técnica de ese requisito aún no se ha resuelto. En el Grupo de Trabajo sobre Preferencias en materia de IA del Internet Engineering Task Force se está elaborando un vocabulario común para las preferencias de uso relacionadas con la IA, trabajo que seguía en curso a mediados de 2026; los borradores de trabajo siguen en fase de elaboración y solo definen un pequeño número de categorías de uso (en esencia, el entrenamiento de la IA y los

[143] El filtrado en la fase de entrada puede consistir en excluir las obras cuyos metadatos indiquen una reserva de derechos, deduplicar el corpus para reducir el riesgo de memorización y excluir contenidos de fuentes que se sabe que carecen de licencia. El filtrado en la fase de generación de resultados puede consistir en detectar y bloquear los resultados que se correspondan estrechamente con obras protegidas identificadas. La importancia relativa que se concede al filtrado en la fase de entrada y en la fase de salida varía según el desarrollador y el tipo de modelo.

[144] Administración del Ciberespacio de China, Operación especial Qinglang para rectificar el caos en las aplicaciones de IA; *Tecnología de ciberseguridad. Requisitos básicos de seguridad para los servicios de inteligencia artificial generativa*, norma nacional recomendada, en vigor desde el 1 de noviembre de 2025, artículo 4.2.1 (filtrado del contenido de los datos de entrenamiento).

[145] Sobre la resolución LAION II y la proliferación de normas de exclusión voluntaria, véanse la parte II, sección B.5, y las referencias que allí figuran al comentario de Keller.

usos relacionados con la búsqueda, con una categoría para la generación aumentada por recuperación que está siendo objeto de debate), y aún no se ha adoptado ninguna norma.[146] Una cuestión conexa que podría merecer la pena analizar es cómo se compagina la obligación de exclusión voluntaria con la prohibición de formalidades para la existencia o el ejercicio de los derechos protegidos por el Convenio de Berna. La respuesta puede depender de la configuración concreta y del efecto jurídico del mecanismo de exclusión voluntaria.

Han surgido en paralelo varias normas que compiten entre sí, y que se dividen en dos grupos fáciles de confundir. La primera expresa únicamente una reserva de derechos, es decir, una exclusión voluntaria. Comprende el TDM Reservation Protocol (TDMRep) y TDM·AI, ambos destinados específicamente a permitir exclusiones voluntarias conformes con el artículo 4, apartado 3, de la Directiva sobre los derechos de autor en el mercado único digital; el mecanismo de señales de contenido de Cloudflare, una extensión patentada del Protocolo de exclusión de robots (robots.txt); e iniciativas más amplias como C2PA e IPTC PLUS, que ofrecen funciones de exclusión voluntaria mediante sus propios vocabularios. Las señales de este tipo ya se utilizan en la práctica frente a los rastreadores de los motores de búsqueda conversacionales basados en IA: según la información disponible, una proporción considerable de los principales sitios de noticias bloquea el rastreador utilizado por uno de esos motores.[147] El segundo grupo va más allá y permite a los titulares de derechos incorporar condiciones de concesión de licencias afirmativas y en formato legible por máquina en esa misma infraestructura web. Por consiguiente, la cuestión ya no es solo si una obra puede utilizarse, sino en qué condiciones (incluido el precio) y la elección entre el pago por rastreo y el pago por inferencia. La norma Really Simple Licensing (RSL), una iniciativa respaldada por editoriales en línea y redes de distribución de contenidos estadounidenses, es un ejemplo destacado; otro sería el modelo de pago por rastreo de Cloudflare, que opera en el plano de la propia solicitud de acceso, devolviendo una respuesta de pago obligatorio a los rastreadores sin licencia. En este segundo grupo, la señal que encuentra un agente automatizado funciona no solo como una exclusión voluntaria, sino como una oferta de licencia. La introducción de agentes de IA podría facilitar la concesión automatizada de licencias en este ámbito.

El cumplimiento y las resoluciones judiciales se complican por la coexistencia de estos dos tipos de reservas en un mismo debate sobre normas que aún no está resuelto. Un mismo proveedor puede operar en ambos, como hace Cloudflare mediante sus mecanismos de señales de contenido y de pago por rastreo, de modo que no siempre es evidente a simple vista si una señal determinada es una mera reserva o una oferta de licencia. La Comisión Europea ha puesto en marcha una consulta sobre los protocolos para reservar derechos a la minería de textos y datos en virtud de la Ley de IA y el Código de Buenas Prácticas de la GPAI, con vistas a elaborar una lista de protocolos legibles por máquina generalmente aceptados que los proveedores de modelos de IA de uso general deben observar para cumplir sus obligaciones con arreglo al Reglamento de IA.[148]

La fragmentación del panorama de las normas es en sí misma un desafío para la observancia. La resolución del Tribunal de Apelación de Hamburgo indica que el formato legible por máquina de una exclusión voluntaria debe evaluarse en función de las capacidades tecnológicas

---

[146] Sobre el Grupo de Trabajo sobre Preferencias en materia de IA del Internet Engineering Task Force, véase https://datatracker.ietf.org/wg/aipref/about/. El TDM Reservation Protocol (TDMRep), TDM·AI, el protocolo Really Simple Licensing (RSL), de RSL Collective, el mecanismo contentsignals.org de Cloudflare, C2PA e IPTC PLUS son algunas de las iniciativas más destacadas que operan en este mismo ámbito.

[147] Un estudio publicado en 2026 constató que muchos de los principales sitios de noticias de los Estados Unidos de América y el Reino Unido bloquean PerplexityBot, el rastreador de Perplexity, en sus archivos robots.txt. Véase BuzzStream, “Most Major News Publishers Block AI Training & Retrieval Bots” (17 de diciembre de 2025, publicado en enero de 2026) <https://www.buzzstream.com/blog/publishers-block-ai-study/>, consultado el 3 de julio de 2026.

[148] Comisión Europea, Consulta sobre los protocolos para reservar derechos a la minería de textos y datos en virtud de la Ley de IA y el Código de Buenas Prácticas de la GPAI (2025); véase https://digital-strategy.ec.europa.eu/es.

existentes en el momento del uso controvertido; pero en un panorama en el que coexisten múltiples normas incompatibles, no existe una respuesta sencilla a la cuestión práctica de qué normas debe observar un desarrollador para cumplir la ley. El riesgo señalado en los comentarios sobre la sentencia es que, a falta de normas compartidas y de gobernanza abierta, las exclusiones voluntarias legibles por máquina puedan convertirse en la base de una nueva capa de intermediación que normalice el acceso a la web abierta mediante pago por rastreo o pago por clic. Que ese riesgo se materialice o no dependerá de las decisiones que se adopten en los organismos de normalización y en la formulación de políticas en los próximos años.

El mecanismo de exclusión voluntaria forma parte de una infraestructura de gestión de derechos con múltiples capas. La primera capa consiste en los identificadores y los metadatos. Esta contiene los identificadores internacionales normalizados consolidados para las obras y los colaboradores, y metadatos legibles por máquina de gestión de derechos y de procedencia que pueden vincular una obra con la información sobre su situación jurídica. Se está trabajando en la recopilación de buenas prácticas en materia de datos de derecho de autor interoperables, fiables y legibles por máquina adecuados al entorno de la IA, en particular en el Equipo Técnico de Infraestructura de Derecho de Autor convocado por el Ministerio de Educación y Cultura de Finlandia,[149] partiendo de la premisa de que una identificación fiable es requisito previo tanto para expresar las preferencias de uso como para rastrear los usos *a posteriori*. Esta capa se sustenta en las disposiciones del WCT, el WPPT y el Tratado de Beijing mencionadas anteriormente sobre la información sobre la gestión de derechos.

La segunda capa consiste en el control de acceso, que comprende las interfaces de programación de aplicaciones y las medidas contra el rastreo mediante las cuales los titulares de derechos y las plataformas tratan de garantizar que solo los agentes autorizados accedan a sus contenidos, junto con las medidas tecnológicas de protección que se utilizan en los sectores de la transmisión por flujo continuo y la edición contra la extracción no autorizada. Esta capa se sustenta en las disposiciones del WCT, el WPPT y el Tratado de Beijing mencionadas anteriormente sobre tecnologías contra la elusión.

La tercera capa es la de la atribución, entendida como la evaluación de la contribución de determinadas obras a un producto concreto con el fin de asignar una remuneración cuando proceda. La atribución en este sentido debe distinguirse de la detección. Las tecnologías de detección tratan de determinar, con cierto grado de probabilidad, si un archivo ha sido generado por una persona o por una máquina; no identifican las obras de origen y no sirven, por sí solas, de fundamento para una reclamación por derecho de autor. Las tecnologías de atribución, en cambio, tratan de estimar la participación de las obras identificadas en un producto generado y, por ese motivo, dependen de cierto grado de cooperación y transparencia por parte del desarrollador. Ya han empezado las inversiones en esta capacidad: un proveedor de tecnología de atribución, que ya participa en el sistema sueco de acuerdos de licencia colectivos mencionado en la parte II, fue adquirido por una gran compañía discográfica en 2026.[150] La capa de la atribución está directamente vinculada a los modelos de remuneración en el plano de los productos mencionados en la parte II, sección B.6. Algunos de esos modelos no se basan en la infracción en la fase de generación de resultados, sino en el efecto sustitutivo de los resultados de determinados modelos.

Una familia distinta de herramientas de procedencia y de marcado de agua, junto con las iniciativas para el etiquetado de los contenidos creados por personas, se ocupa de la

[149] El Equipo Técnico de Infraestructura de Derecho de Autor, convocado bajo los auspicios del Ministerio de Educación y Cultura de Finlandia, publicó en 2025 un primer informe titulado “Datos de derecho de autor interoperables, fiables y legibles por máquina en la era de la IA”.

[150] Warner Music Group adquirió el proveedor de tecnología de atribución Sureel en junio de 2026. La tecnología de ese mismo proveedor se utiliza en el acuerdo de licencia en el que participa la STIM, examinado en la parte II, sección B.3.

autenticidad y el origen de los productos, y no del uso de obras protegidas en el entrenamiento; esas herramientas atañen principalmente a las cuestiones relativas a los productos que el presente estudio no aborda, y se mencionan aquí únicamente para situar la capa de la atribución dentro de la infraestructura más amplia.

3. La jurisprudencia en su conjunto

El análisis del conjunto de la jurisprudencia examinada permite formular varias observaciones transversales sobre la actual observancia del derecho de autor en el contexto del entrenamiento de la IA generativa.

La primera es que los tribunales han actuado, en general, con cierta cautela y, por lo común, se han abstenido de dictar resoluciones que pretendan resolver de forma exhaustiva el panorama doctrinal. Las resoluciones dictadas hasta la fecha han abordado tecnologías concretas (la difusión latente en el caso *Getty*, los modelos de lenguaje en el caso *GEMA* y los libros adquiridos de forma lícita e ilícita en el caso *Bartz*), situaciones procesales concretas (la moción de desestimación en el litigio del *New York Times* y la sentencia por vía sumaria en los casos *Bartz* y *Kadrey*) y cuestiones concretas (el modelo como copia en los casos *GEMA* y *Getty* y el uso leal en los casos *Bartz* y *Kadrey*, entre otras). Aún no han dado lugar a un marco completo en ningún ordenamiento jurídico. La posición alemana, aunque ya ha quedado reflejada en tres resoluciones, no es uniforme en sí misma: la línea LAION acepta la aplicación de la excepción relativa a la minería de textos y datos al entrenamiento, mientras que la línea GEMA adopta un planteamiento diferente. En los Estados Unidos de América la situación también está estratificada y a la espera de que se pronuncien los tribunales de apelación y, en última instancia, el Tribunal Supremo de los Estados Unidos de América en una o varias resoluciones.

La segunda es que la jurisprudencia ha dado lugar, no obstante, a algunos puntos de convergencia. Existe consenso en cuanto a la exigencia, recogida en el Convenio de Berna, de un autor humano para generar la originalidad necesaria a efectos de la protección de los resultados de la IA. Este puede observarse en los Estados Unidos de América, Kenya,[151] Alemania en primera instancia y la Unión Europea, donde el criterio de originalidad reafirmado por el TJUE en el asunto *Mio*, relativo a la protección por derecho de autor de muebles, presupone un autor humano.[152] También existe convergencia sobre la afectación del derecho de reproducción por la constitución de conjuntos de datos de entrenamiento (en Alemania, China y el Canadá). Por último, hay también cierta convergencia en cuanto a la pertinencia de la sustitución mediante resultados para el análisis (en los Estados Unidos de América y, en perspectiva, en la Unión Europea, mediante la aplicación de la regla de los tres pasos). Entre las cuestiones sobre las que la jurisprudencia sigue mostrándose dividida se encuentran si los modelos entrenados contienen copias de las obras utilizadas para el entrenamiento, el reentrenamiento y el ajuste fino; si la excepción relativa a la minería de textos y datos se

[151] *Aryeh Movement Limited v. Cynthia Beldina Akoth Okello*, COPTA/E001/2025 (Tribunal de Derecho de Autor de Kenya) (en la que se resuelve que el artículo 22.3 de la Ley de Derecho de Autor (Cap. 130) presupone la intervención humana en el despliegue de un “esfuerzo suficiente”, de modo que una obra generada de forma autónoma por un sistema de IA no puede alcanzar el umbral legal a menos que una intervención humana suficiente le confiera originalidad).

[152] *Mio and Others*, C-580/23, agrupado con *USM v. konektra*, C-795/23 (TJUE, sentencia de 4 de diciembre de 2025). Los asuntos agrupados se referían a la protección del derecho de autor de obras de arte aplicado, concretamente de muebles (en el asunto C-580/23, las mesas de comedor del Palais Royal). El Tribunal reafirmó que un objeto solo puede considerarse obra protegida si es original, en el sentido de que constituye una creación intelectual propia de su autor que refleja decisiones libres y creativas que expresan su personalidad, y que las obras de arte aplicado están sujetas al mismo requisito de originalidad que las demás obras. La sentencia no se refería a la inteligencia artificial; la exigencia de un autor humano para la protección de los resultados de la IA se deriva de este criterio de originalidad, y no de ningún pronunciamiento del Tribunal dirigido específicamente a la IA. Además, los autores pueden enfrentarse a un nuevo obstáculo en los procedimientos de observancia, a saber, demostrar que su obra no ha sido creada por IA o, al menos, que su contribución ha sido suficiente para justificar la protección por derecho de autor.

extiende al entrenamiento de la IA y, en caso afirmativo, de qué manera; y cuál es el alcance de la jurisdicción transfronteriza.

La tercera observación es que los desafíos prácticos señalados en las secciones C.1 y C.2, más arriba, a saber, la transparencia, la proliferación de normas de exclusión voluntaria y las asimetrías de costos e información entre el desarrollador y el titular de los derechos, probablemente condicionarán la evolución de la jurisprudencia en los próximos meses y años. Aun cuando los tribunales resuelvan las cuestiones sustantivas de derecho de autor, la posibilidad práctica de hacer valer las pretensiones de los titulares de derechos dependerá de los mecanismos de que estos dispongan para averiguar el uso de sus obras, expresar exclusiones voluntarias y presentar pruebas en una forma con la que los tribunales puedan trabajar. En este ámbito, el diseño técnico y el análisis jurídico están estrechamente relacionados, y la observancia dependerá de las decisiones que se adopten en ambos campos.

CONCLUSIÓN

En el presente estudio se constata una evolución rápida y desigual en dos esferas: hay jurisdicciones que están experimentando distintas combinaciones de excepciones, mecanismos de concesión de licencias y regímenes de remuneración, y un conjunto de jurisprudencia que ha arrojado luz sobre las cuestiones, pero que aún no ha dado lugar a un marco estable. Los Estados miembros que actualmente están estudiando respuestas legislativas o reglamentarias lo hacen en un contexto más desarrollado que el que existía hace tan solo 12 meses; los puntos en los que la experiencia y la jurisprudencia han convergido ofrecen una base parcial, mientras que los puntos en los que siguen divididas exigirán decisiones de política que el presente estudio ha tratado de exponer, pero no de resolver. Esas decisiones seguirán adoptándose en el plano nacional y, cuando los Estados miembros así lo decidan, se debatirán en el plano internacional en el seno del Comité Permanente de Derecho de Autor y Derechos Conexos y del marco más amplio de la OMPI.

El estudio pone de manifiesto que cada Estado miembro puede alcanzar de forma diferente el equilibrio entre los intereses de los titulares de derechos y los de los desarrolladores de IA. Del análisis se desprenden otras cinco preguntas más concretas.

La primera es si las reproducciones realizadas al crear un corpus y realizar el entrenamiento de un modelo deben requerir autorización o si entran dentro de una excepción o limitación, ya sea específica de la IA, una disposición sobre minería de textos y datos o una cláusula general, y en qué condiciones, entre otras, el acceso lícito, la finalidad del entrenamiento y el carácter comercial o no comercial del uso.

La segunda es el diseño de cualquier mecanismo de reserva donde se apliquen. Ello incluye la forma de exclusión voluntaria, teniendo en cuenta que las reservas legibles por máquina presuponen una capacidad de la que los autores y los artistas intérpretes o ejecutantes individuales pueden no disponer. Por el contrario, un régimen de reservas que no sea automatizable a gran escala no ofrece a los desarrolladores una vía clara para garantizar la observancia. También debería examinarse la compatibilidad de un régimen de exclusión voluntaria con el artículo 5.2) del Convenio de Berna.

La tercera es si la compensación debe estar vinculada a la ingestión, a la sustitución por los resultados, o a ambos; si debe adoptar la forma de licencias de mercado basadas en derechos exclusivos o de un derecho legal a remuneración; y, en cualquiera de los dos casos, a través de qué instituciones debe canalizarse. Esto incluye si los mandatos de gestión colectiva se extienden a los usos relacionados con el entrenamiento, si se necesitan formas ampliadas o nuevas de concesión de licencias colectivas, y cómo se pueden diseñar las cadenas

contractuales para que el valor llegue a los autores y los artistas intérpretes o ejecutantes cuyas obras se utilizan.

La cuarta es la transparencia: qué requisitos de registro y divulgación del material de entrenamiento son necesarios para permitir el ejercicio de los derechos y la gestión de la remuneración, y cómo pueden conciliarse esas obligaciones con los secretos comerciales legítimos y las preocupaciones en materia de competencia.

La quinta y última es si las medidas de reparación siguen siendo eficaces y proporcionadas cuando resulta técnicamente inviable eliminar las obras de un modelo entrenado, y cómo debe ejercerse la competencia jurisdiccional cuando el entrenamiento se lleva a cabo en el extranjero, pero los servicios resultantes se ofrecen en el territorio nacional.

[Fin del documento]