IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

https://idp.uoc.edu

Derechos de autor y entrenamiento de sistemas de IA generativos: las obligaciones de transparencia y la minería de textos y datos en la normativa europea

Jorge Luis Ordelin Font
Centro de Investigación y Docencia Económicas (CIDE), México

Fecha de presentación: agosto 2024
Fecha de aceptación: enero 2025
Fecha de publicación: marzo 2025

Resumen

La relación de los derechos de autor y la IA, en particular la generativa, supone innumerables retos y contradicciones. En el centro de todos los debates, se encuentran los derechos de los titulares de las obras y demás prestaciones artísticas. Las obras y prestaciones son utilizadas para entrenar modelos de IA, como fuente de datos, lo cual supone una pérdida económica para los titulares. El Reglamento de IA de la Unión Europea ofrece algunas herramientas para poder contrarrestar esta situación, sin embargo, su alcance es limitado. El presente artículo tiene como objetivo analizar los retos y limitaciones de estas herramientas, en particular la transparencia algorítmica y la limitación de la minería de textos y datos.

Palabras clave

derechos de autor; entrenamiento de datos; IA generativa; minería de textos y datos; transparencia algorítmica

IDP REVISTA D'INTERNET, DRET I POLÍTICA
REVISTA DE INTERNET, DERECHO Y POLÍTICA

https://idp.uoc.edu

Copyright and training of generative AI systems: transparency obligations and text and data mining in European regulations

Abstract

The relationship between copyright and AI, particularly generative, poses countless challenges and contradictions. The rights of the owners of the works and other artistic performances are at the heart of all discussions. The works and said performances are used to train AI models as a data source, which is an economic loss for the owners. The EU AI Regulation provides some tools to counteract this situation. However, its scope is limited. This article discusses the challenges and limitations of these tools, particularly algorithmic transparency and the limitation of text and data mining.

Keywords

copyright; data training; generative AI; text and data mining; algorithmic transparency

Introducción

La intersección entre los derechos de autor y la inteligencia artificial (IA), particularmente la IA generativa,[1] genera múltiples áreas de interés. Al respecto, la Recomendación sobre la Ética de la IA de la UNESCO, ha llamado a «promover investigaciones sobre la intersección entre la IA y la propiedad intelectual», en particular, analizar si se puede «proteger con derechos de propiedad intelectual las obras creadas mediante tecnologías de la IA y la manera de hacerlo», y «cómo afectan las tecnologías de la IA a los derechos o los intereses de los titulares de derechos de propiedad intelectual cuyas obras se utilizan para investigar, desarrollar, entrenar o implantar aplicaciones de IA» (UNESCO, 2021).

La Resolución del Parlamento Europeo de 20 de octubre de 2020, al abordar la relación de los derechos de propiedad intelectual para el desarrollo de las tecnologías relativas a la IA resaltó la necesidad de que la titularidad de los derechos de autor se asigne únicamente a las personas creadoras, solo si el titular concede la autorización de uso de las obras, excepto, cuando apliquen excepciones (Parlamento Europeo, 2020).

El 13 de junio de 2024 fue aprobado el Reglamento (UE) 2024/1689 del Parlamento Europeo y del Consejo por el que se establecen normas armonizadas en materia de inteligencia artificial.[2] Con su adopción se esperaban soluciones a algunos de los problemas jurídicos planteados a partir de la intersección entre derechos de autor e IA generativa, sin embargo, lo cierto es que la norma comunitaria no cumple con este rol, al menos de manera expresa. El European Innovation Council y la SMES Executive Agency, tras la aprobación del Reglamento de IA, dividieron en dos categorías principales los problemas que se plantean entre los derechos de autor y la IA generativa:

1) la posible infracción de los derechos de autor por parte de los desarrolladores de estas herramientas por el uso de materiales protegidos por derechos de autor para entrenar sus algoritmos y;

2) si las obras producidas por o con herramientas de IA están protegidas por los derechos de autor y a quién le corresponden los derechos de autor resultantes[3] (European Innovation Council y la SMES Executive Agency, 2024). Sin

1. Constituye una subrama dentro de la IA que hace referencia a aquellos modelos de IA que generan contenidos a partir de entradas de usuarios. Los contenidos incluyen imágenes, vídeos, texto y audio. Algunos de estos modelos pueden ser unimodales, generar un solo tipo de contenido o multimodales, en los cuales se generan contenidos de múltiples modos, por ejemplo texto, imágenes y vídeo. JONES, E. (2023). A efectos de este trabajo se utilizará el concepto de contenidos sintéticos, para hacer referencia a toda la información, dígase imágenes, videos, clips de audio y texto que ha sido generado en su totalidad por la IA Generativa. Si bien este concepto se utiliza en el Reglamento de IA en el considerando 133 y artículo 50.2, no está definido en la norma, sin embargo, del contexto se puede colegir su alcance.
2. 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. (en adelante Reglamento de IA).
3. Aun cuando no se reconozcan derechos de autor sobre los contenidos sintéticos, lo cierto es que, estos contenidos son comercializados en múltiples plataformas digitales. El impacto de la IA no es uniforme para cada sector cultural, en particular en relación con los perjuicios económicos que se pueden producir a partir de la pérdida de ingresos o reducción de la capacidad de obtener beneficios económicos del trabajo. Sobre esta distinción, (Hutiri, Papakyriakopoulos y Xiang, 2024).

embargo, existen diversas dudas sobre cómo solucionar dichos problemas debido a que la regulación no brinda respuestas satisfactorias a estos.

El presente artículo parte de reconocer las limitaciones que son intrínsecas a la norma. El Reglamento de IA no tiene entre sus objetivos regular el tema de la IA generativa y los derechos de autor, a su vez, el sistema normativo de derechos de autor al que hace referencia tampoco ofrece soluciones jurídicas a los retos que esta tecnología plantea. Por ello, refiriéndose al Reglamento de IA algunos autores consideran que es una «escueta regulación indirecta» (Jiménez Serranía, 2024) o «esperanza limitada en beneficio de los titulares de derechos y de los desarrolladores» (Mezei, 2024). Sin embargo, establece mecanismos o herramientas que bien podrían ser utilizados en la protección de los derechos de autor y conexos ante el desarrollo y uso de esta tecnología. A partir de este contexto el objetivo de este artículo es analizar los retos y limitaciones de los mecanismos de transparencia algorítmica y minería de textos y datos (TDM) que se reconocen en el Reglamento de IA para la protección de los derechos de autor ante el desarrollo de modelos de IA de uso general y, en particular, los modelos de IA generativa.[4]

1. La obligación de transparencia algorítmica y datos de entrenamiento

La transparencia algorítmica constituye un principio de la IA, supone la capacidad de los diferentes actores del ciclo de vida de la IA de proporcionar información significativa sobre el modelo de IA, e incluye, entre otros aspectos, información sobre las fuentes de datos/insumos, factores, así como aquellos procesos que conducen al resultado, con el fin de comprenderlo (OCDE, 2024). Desde el punto de vista de la generación de contenido sintético y desde los derechos de autor se materializa en dos perspectivas. La primera, saber con qué datos han sido entrenados los sistemas de IA y, la segunda, conocer si un determinado contenido ha sido creado o modificado con IA. En relación con la primera de estas perspectivas es imprescindible poder determinar cuáles son los datos que han sido utilizados en el entrenamiento, sus fuentes y la legalidad de su utilización.[5]

El uso de obras por parte de la IA no representa un riesgo propiamente dicho, al menos como lo ha entendido y conceptualizado el Reglamento de IA. Sin embargo, la licitud en el uso de los datos utilizados en dicho entrenamiento, dentro de los cuales encontramos aquellos cuyas fuentes están protegidas por los derechos de autor no es solo un tema de transparencia algorítmica, sino también, económico, que afecta a los autores y a los titulares de derechos conexos. Los modelos de IA generativa «no se inspiran» en las obras que se utilizan para su entrenamiento, más que inspiración, dichas obras sirven como ejemplos para generar nuevos «datos sintéticos» similares a aquellos que fueron utilizados en su entrenamiento.[6] Esto implica que, para lograr una mayor calidad de sus resultados, se necesiten mayores y variados datos de entrenamiento,[7] en cantidades verdaderamente astronómicas, cuyo acceso en la mayoría de los supuestos es realizado sin el consentimiento de sus titulares.[8]

En el Reglamento de IA, los proveedores de modelos de IA de uso general tienen obligaciones de transparencia en relación con los datos utilizados para el entrenamiento, la identificación de cómo se obtuvieron y seleccionaron

4. Los modelos de IA generativa son un ejemplo de modelos de IA de uso general. En este trabajo se hace referencia específicamente a los modelos de IA generativa.
5. Esto no es sencillo de probar, en la actualidad se habla de una crisis para demostrar la infracción de los derechos de autor en relación con este tipo de tecnología. (Crawford y Schultz, 2024).
6. El uso de obras en el entrenamiento de los modelos está relacionado con la calidad de los modelos generativos particularmente. Cuando los directivos de OpenAI comparecieron ante la Cámara de los Lores en Reino Unido reconocieron que limitar los datos de entrenamientos a obras que se encuentren en dominio público, no proporcionaría sistemas de IA que satisfagan las necesidades de los ciudadanos de hoy. OpenAI (2023).
7. Actualmente existen otras teorías que hacen referencia a la necesidad de contar con datos que reflejen interacciones de tipo humano, en vez de calidad o diversidad. Sin embargo, no siempre es posible decir cuáles datos constituyen una demostración refleja de estilos humanos. (Shen, 2024).
8. La herramienta de generación de video por IA Júpiter (Gen-3) creada por la empresa Runway se entrenó recopilando miles de videos de Youtube y películas pirateadas (Cole, 2024). Udio reconoció la copia de las obras de los demandantes como parte del desarrollo de una nueva tecnología, cuyos resultados finales en sí mismo no son infractores. Respuesta a la demanda de UMG Recordings, Inc. v. Uncharted Labs, Inc. 2024.

dichos datos, sus fuentes, entre otras obligaciones.[9] Estos actores tienen una función y responsabilidad particular a lo largo de la cadena de valor de la IA y, en correspondencia con ello, se han establecido medidas de transparencia proporcionales, entre las que se encuentran la elaboración de la documentación, su actualización y facilitación a los proveedores posteriores.[10] La obligación de transparentar los datos utilizados en el entrenamiento incluye aquellos textos y datos protegidos por los derechos de autor. En un intento de lograr dicha protección, se ha establecido la obligación 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.[11] Sin embargo, la configuración de la obligación no es clara, en particular, en relación con su alcance y efectos.

El resumen debe ser público, pero al propio tiempo proteger los secretos comerciales y la información empresarial confidencial;[12] debe ser «exahustivo en general en su alcance en vez de técnicamente detallado»; sencillo y eficaz y, al propio tiempo, ser proporcionado en forma descriptiva;[13] facilitar a las partes con intereses legítimos, incluidos los titulares de derechos de autor que ejerzan y hagan cumplir sus derechos, aunque no existe la obligación de verificar ni proceder a «una evaluación obra por obra de los datos de entrenamiento en cuanto al respeto de los derechos de autor».[14] En otras palabras, debe ser muchas cosas, pero parece poco efectivo para defender los derechos de titulares concretos y, especialmente, hacer viable la reserva de derechos. Los derechos de autor continúan siendo derechos individuales aun cuando el uso por la IA sea masivo, en otras palabras, si no se tiene en cuenta el uso individual de cada obra, es muy poco probable la viabilidad de la protección.

Al propio tiempo, el cumplimiento de la obligación de transparencia deberá ser proporcionado y adecuado al tipo de proveedor de modelos,[15] esto significa que quienes desarrollan o utilizan modelos con fines no profesionales o de investigación científica no deban cumplir con dicha obligación, aunque se anima a su cumplimiento voluntario. También se debe tener en cuenta el tamaño del proveedor, así como que su cumplimiento no suponga un coste excesivo para las pymes, incluyendo las empresas emergentes, ni desincentive su uso. Además, existen adecuaciones si se trata de la modificación o ajuste del modelo, en cuyo caso la obligación de transparencia está limitada a dichas modificaciones o ajustes, incluyendo la declaración de nuevas fuentes de datos de entrenamiento.

En materia de derechos de autor, el cumplimiento de las obligaciones de transparencia establecidas no es del todo claro, el segundo borrador del *Código de buenas prácticas de la IA para fines generales* tampoco ofrece luces sobre cómo estas «adecuaciones» tendrían lugar, en particular al momento de interpretar lo que pudiera ser un «coste excesivo» para las pymes, que no «desincentive» el uso de estos modelos, y cuánto cambiaría dicha obligación en estos casos, ¿ante quién se podría alegar la excepción de su cumplimiento y cómo?, ¿cuáles serían las diferencias en el cumplimiento de esta obligación entre aquellos «grandes» proveedores de los modelos y las pequeñas empresas? Son muchas las dudas que existen.

2. El uso lícito de obras como fuente de datos de entrenamiento: la excepción de la minería de textos y datos

El cumplimiento efectivo de la obligación de transparencia es esencial para determinar si el uso de las obras puede ser justificado bajo el régimen de la minería de textos y datos (TDM). Tras la adopción del Reglamento de IA se ha considerado como lícito el uso de obras en el entrenamiento de modelos generales de IA siempre y cuando se utilicen técnicas de TDM. Sin embargo, aun cuando se

9. Art. 53.1.a) yb) en relación con los Anexos XI y XII.
10. Considerando 101 del Reglamento de IA.
11. Considerando 107 em relación com el artículo 53.1.d) del Reglamento de IA.
12. Considerando 107 del Reglamento de IA.
13. Se reconoce, por ejemplo, enumerar los principales conjuntos o recopilaciones de datos que hayan servido para entrenar al modelo, como los grandes archivos de datos o bases de datos privados o públicos, y así como proporcionar una explicación descriptiva sobre otras fuentes de datos utilizadas. Considerando 107 del Reglamento de IA.
14. Considerando 108 del Reglamento de IA.
15. Considerando 108 del Reglamento de IA.

ha contemplado que dicha actividad es permitida en el marco comunitario de protección de los derechos de autor;[16] lo cierto es que este no es un tema resuelto, ya sea porque se considera que la excepción legal no coincide con el entrenamiento de los modelos de IA generales o, porque la figura es insuficiente para proteger a los titulares de derechos.[17]

2.1. La excepción de TDM y el Reglamento de IA

El Reglamento de IA reconoce que la IA supone un desafío para los artistas, autores y demás creadores, particularmente por la manera en que se crea, distribuye, utiliza y consume el contenido creativo. En este sentido, establece como regla general que cualquier uso de contenidos protegidos por derechos de autor requiere la autorización del titular de los derechos de que se trate, es decir, no se afecte el cumplimiento de las normas previstas en el Derecho de la Unión,[18] salvo que se apliquen las excepciones y limitaciones pertinentes en materia de derechos de autor, en particular, las establecidas en la Directiva (UE) 2019/790 sobre el mercado único digital, que permiten reproducciones y extracciones de obras y otras prestaciones con fines de prospección de textos y datos en determinadas circunstancias.

Los proveedores que introduzcan modelos de IA de uso general en el mercado deben adoptar directrices para el cumplimiento del Derecho de la Unión en materia de derechos de autor y derechos afines,[19] en particular, para detectar y cumplir la reserva de derechos expresada por los titulares con arreglo al artículo 4, apartado 3, de la Directiva (UE) 2019/790,[20] siendo el cumplimiento de esta obligación independiente de la jurisdicción en la que tengan lugar los actos relacionados con el entrenamiento de los modelos de IA de uso general.[21] Además, debe tenerse en cuenta que esta obligación referida expresamente en el Reglamento de IA es ejemplificativa, lo que significa que, los proveedores deban cumplir otras obligaciones, no solo en relación con la excepción de TDM sino, de manera general, cualquier otra referida a los derechos de autor en el marco europeo.[22]

En el contexto de la TDM aplicada a los modelos de IA generales se ha vuelto medular:

> «[D]eterminar si es posible aplicar la limitación de la minería cuando lo que se pretende no es tanto obtener una información concreta del análisis de los macrodatos, sino entrenar un sistema de inteligencia artificial generativa» (García Vidal, 2024, pág. 3).

Las posiciones al respecto no son uniformes, por un lado se encuentran quienes consideran que la formulación actual de la TDM permite el entrenamiento de sistemas de IA generativos, mientras que otros autores, consideramos

16. Esta posibilidad se preveía desde el párrafo 17 de la Resolución del Parlamento Europeo de 20 de octubre de 2020, al expresarse «debe evaluarse a la luz de las normas existentes sobre las limitaciones y excepciones a la protección mediante derechos de autor, como la prospección de texto y datos». Esta posición ha sido confirmada tras la aprobación del Reglamento de IA por el *European Innovation Council* y la *SMEs Executive Agency* (2024).
17. En la consulta abierta sobre Derechos de Autor e Inteligencia artificial del Gobierno de Reino Unido se plantean varias posiciones sobre el tema: 1) mantener la regulación tal como está; 2) reforzar los derechos de autor exigiendo licencias en todos los casos; 3) establecer una excepción amplia que permitiría la extracción de datos de obras protegidas por derechos de autor, incluida la formación de IA, sin el permiso de los titulares de los derechos, 4) una excepción a la extracción de datos que permita a los titulares reservarse sus derechos. Esta última posición es similar a la de la Unión Europea aunque a diferencia incluye la explotación con fines comerciales. Intellectual Property Office, Department For Science, Innovation & Technology And Department For Culture, Media & Sport (2024).
18. Considerando 108 del Reglamento de IA
19. Las obligaciones referidas a los derechos de autor también deberán ser cumplidas por los proveedores de modelos de IA que los divulgen con arreglo a una licencia libre y de código abierto. Art. 53.2 del Reglamento de IA.
20. Considerando 106 en relación con el Art. 53.1. c) del Reglamento de IA.
21. Considerando 106 del Reglamento de IA.
22. Ni el Reglamento de IA ni los borradores del Código de buenas prácticas de la IA para fines generales ha determinado con claridad cuáles son las obligaciones que tienen los proveedores de modelos de IA generales en materia de derechos de autor conforme al ordenamiento europeo. En este sentido solo se propone, entre otras medidas, que cada proveedor establezca una política interna para cumplir con la legislación, aplicable a todas las fases del desarrollo de un modelo de IA de propósito general, incluida la recopilación de datos, la formación, las pruebas y la comercialización, pero no qué debe contener dicha política. (EU AI Office, 2024).

que no es posible, a partir de una interpretación teleológica de la norma y de la tecnología.[23]

2.2. TDM e IA generativa

Para poder determinar la excepción de TDM permite el entrenamiento de modelos de IA generativos es importante distinguir entre la finalidad de las técnicas de TDM y la finalidad de la limitación y excepción. La Directiva (UE) 2019/790 entiende como TDM, «toda técnica analítica automatizada destinada a analizar textos y datos en formato digital a fin de generar información que incluye, sin carácter exhaustivo, pautas, tendencias o correlaciones»,[24] siendo su finalidad «el tratamiento de grandes cantidades de información con el fin de adquirir nuevos conocimientos y descubrir nuevas tendencias, beneficiando a la comunidad de investigadores y apoyando a la innovación».[25]

El legislador europeo configuró jurídicamente la excepción a partir de la finalidad de la técnica analítica que, a su vez, permite el ejercicio de otros derechos. La finalidad expresa de la excepción no es un derecho humano, como suele suceder en materia de limitaciones y excepciones de derechos de autor. Ello se debe, entre otras razones, a que el uso de la tecnología puede ser diverso, como son la adopción de decisiones empresariales complejas, prestar servicios públicos y el desarrollo de nuevas aplicaciones o tecnologías,[26] dentro de las que se pudieran incluir los modelos de IA de uso general. De hecho, la finalidad de la excepción prevista en el artículo 4 de la Directiva (UE) 2019/790 es la de alentar la innovación en el sector privado.

Sin embargo, no debe confundirse el uso de datos en el entrenamiento de *machine learning* con el entrenamiento de datos en la IA Generativa.[27] Las diferencias no solo se sustentan en las finalidades, sino también, en las técnicas que se utilizan en uno u otro caso. Mientras la TDM está referida al proceso de extraer información útil, identificar relaciones, conocimiento significativo de grandes conjuntos de datos no estructurados, como texto, imágenes, videos, etc., descubrir nueva información y revelar patrones (Sag, 2019);[28] en el caso de los modelos generativos los datos de entrenamiento constituyen la base del proceso de aprendizaje, el modelo aprende de los patrones que subyacen en estos datos, pero también, de las formas de expresión que contienen; estas constituyen el conjunto de ejemplos que le permiten «aprender» a estos modelos . Su finalidad no es extraer la información sino generar nuevos datos similares a los utilizados en el entrenamiento.

Desde un punto de vista tecnológico la TDM es complementaria al entrenamiento,[29] puede ser previa o coexistir con este, pero no es lo mismo. Ello implica que para subsumir el entrenamiento de datos en la excepción de TDM esta debería estar configurada de forma suficientemente amplia para abarcar todo el proceso de entrenamiento de los modelos generativos, incluyendo todos los actos que en relación con las obras se realizan durante esta etapa y con una finalidad distinta a la consagrada actualmente en la Directiva (UE) 2019/790.

Como ya algunos autores habían señalado, incluso antes de la adopción del Reglamento de IA, el ámbito de aplicación de la excepción era limitado, debido a que no cubría todo el espectro de técnicas de TDM (González Otero, 2019). La excepción solo hace referencia a reproducciones y extracciones de obras y prestaciones accesibles de forma legítima, por ende, todas las etapas del entrenamiento

23. La sentencia del Tribunal Regional (Landgericht) de Hamburgo de 27 de septiembre del 2024 considera que sí es posible concebir el entrenamiento de la IA Generativa dentro de la regulación de la excepción en el ordenamiento jurídico alemán y europeo, criterio contrario sostienen Dornis y Stober para quienes una equiparación del entrenamiento de modelos generativos de IA con el TDM «clásico», es tecnológica y conceptualmente errónea (2024, pág. 94).
24. Art.2.2 Directiva (UE) 2019/790.
25. Considerando 8 Directiva (UE) 2019/790.
26. Considerando 18 Directiva (UE) 2019/790.
27. En la IA Generativa adquiere particular importancia los denominados datos multimodales que son el conjunto de datos (imágenes, audio, vídeo) a partir de los cuales se enseña a los modelos a interpretar «imágenes visuales, acciones y secuencias en contextos reales, patrones lingüísticos y matices del habla». OpenAI (2024).
28. Incluye diversas etapas como pueden ser el preprocesamiento de los datos, extracción de características, la aplicación de técnicas específicas en función de los objetivos del proyecto, formación y evaluación de los modelos, interpretación y visualización de los resultados. (Shiksha online, 2023).
29. El proceso de entrenamiento es más complejo, además de la aplicación de la TDM, existe una etapa de recopilación de datos, su preparación, el entrenamiento propiamente dicho y la evaluación y ajuste del modelo.

no siempre quedan contempladas en los actos permitidos en la norma, especialmente cuando se aplican nuevas técnicas en las cuales las obras no son almacenadas, sino que el acceso tiene lugar desde los propios enlaces, lo que constituye un acto de comunicación pública.[30]

Dado que la interpretación y aplicación del régimen de excepciones y limitaciones en materia de derechos de autor es restrictiva, sin que quepa la interpretación analógica,[31] la única posibilidad existente es modificar la regulación del artículo 4 de la Directiva (UE) 2019/790 todo ello con la finalidad de que los actos que se realicen en la aplicación de esta tecnología cumplan con los actos permitidos en la excepción; además del resto de requerimientos legales, el acceso de forma legítima a la obra u prestación, que se haya puesto a disposición del público en línea y, que los titulares no se hayan reservado de forma adecuada los derechos de hacer reproducciones y extracciones con dicha finalidad.[32]

Como afirman Dornis y Stober el Reglamento de IA tiene una relevancia limitada para la interpretación de la «minería de textos y datos» (2024, pág. 96). Sin embargo, uno de los aspectos donde se puede afirmar que ha tenido un mayor impacto es en la importancia que adquiere la reserva de derechos, figura que marca los límites de aplicación de la excepción de TDM. Esta operará siempre y cuando el uso de las obras y prestaciones «no esté reservado expresamente por los titulares de derechos de manera adecuada, como medios de lectura mecánica en el caso del contenido puesto a la disposición del público en línea»,[33] en otras palabras, si la reserva existe será necesario recabar la autorización de los titulares de derechos para el uso de estas obras y prestaciones en actividades de prospección de textos y datos; mientras que, si la reserva no se ha realizado, se podrá aplicar la excepción prevista.[34]

La reserva de derechos incluye la prohibición de todos los procesos automatizados por bots incluyendo el *scrapping* y la descarga automatizada de contenidos. En principio, la Directiva (UE) 2019/790 considera como medios para realizar dicha reserva «la utilización de medios de lectura mecánica, incluidos los metadatos y las condiciones de un sitio web o un servicio», aunque también se declaran en su considerando 18 otros, como son los acuerdos contractuales o una declaración unilateral. Por tanto, la reserva puede ser realizada tanto por medios tecnológicos como jurídicos. Sin embargo, la previsión legal parece trastocarse con las prácticas de las empresas tecnológicas que desarrollan estos modelos, y plantea innumerables preguntas sobre su viabilidad.[35] Aun cuando el borrador del *Código de buenas prácticas de la IA para fines generales*

30. Esta es, por ejemplo, la tecnología que se ha utilizado en el caso LAIO. (Guadamuz, 2023). Los actos concretamente permitidos por la excepción son concretamente los siguientes: a) la reproducción temporal o permanente, total o parcial, por cualquier medio y de cualquier forma de las bases de datos (Artículo 5 a) de la Directiva 96/9/CE); b) la extracción y/o reutilización de la totalidad o de una parte sustancial del contenido de ésta cuando la obtención, la verificación o la presentación de dicho contenido representen una inversión sustancial desde el punto de vista cuantitativo o cualitativo (Art. 7.1 de la Directiva 96/9/CE); c) derecho de reproducción de obras y demás prestaciones (Art. 2 Directiva 2001/29/CE); d) la reproducción total o parcial de un programa de ordenador por cualquier medio y bajo cualquier forma, ya fuere permanente o transitoria; y la traducción, adaptación, arreglo y cualquier otra transformación de un programa de ordenador, así como la reproducción de los resultados de tales actos (Artículo 4, apartado 1, letras a) y b), de la Directiva 2009/24/CE); e) el derecho de reproducción y puesta a disposición al público de las publicaciones de prensa por parte de prestadores de servicios de la sociedad de la información (publicaciones de prensa en línea) (Artículo 15, apartado 1, Directiva (UE) 2019/790).
31. La interpretación que se realiza en la sentencia del Tribunal Regional (Landgericht) de Hamburgo de 27 de septiembre del 2024 es parcial. El juzgador consideró que el uso relevante para los derechos de autor que requiere justificación por la TDM no es la realización de un «análisis de patrones», sino la reproducción de la obra protegida por los derechos de autor. En este último caso, si bien distingue entre la creación de un conjunto de datos, el posterior entrenamiento de la red neuronal artificial con este conjunto de datos y el uso posterior de la IA entrenada para crear nuevos contenidos de imagen, solo analiza la reproducción de la obra en relación con la creación del conjunto de datos, excluyendo el uso de la obra en el resto de las etapas del entrenamiento.
32. Considerando 18 Directiva (UE) 2019/790.
33. Art. 4.3 de la Directiva (UE) 2019/790.
34. Considerando 105 del Reglamento de IA.
35. Sobre cómo se implementaría la reserva de derechos, *vid*. Mezei, P. (2024). Dudas sobre la implementación efectiva de este mecanismo también han sido planteadas en consulta abierta sobre Derechos de Autor e Inteligencia artificial del Gobierno de Reino Unido. (Intellectual Property Office, Department For Science, Innovation & Technology And Department For Culture, Media & Sport, 2024).

propone algunas medidas que permitirán la puesta en práctica del régimen de exclusión,[36] continúa siendo poco viable para la protección efectiva de los derechos de autor en el entrenamiento de datos de la IA Generativa.

Los sistemas de IA utilizan millones de obras y fuentes de datos diversas, no solo protegidas por los derechos de autor, sino también, de cualquier otra naturaleza, en su mayoría extraídas de Internet, a partir de la aplicación de técnicas conocidas como el *web scrapping*[37] y sin pedir autorización por dichos usos. La exclusión voluntaria no opera retroactivamente, funciona bajo el principio de pedir perdón en vez de permiso o autorización, y solo opera en relación con aquellos datos que se utilicen en futuros conjuntos de entrenamiento, debido a que no se eliminarán aquellos que han sido utilizados de forma previa en los modelos.[38] Por tanto serán los autores y titulares de derechos quienes tengan la responsabilidad de adoptar las medidas jurídicas y tecnológicas necesarias para excluir el acceso a sus contenidos, mientras que, las empresas tecnológicas continúan creyendo que todo lo que se encuentra en Internet es público y puede ser libremente utilizado.[39] Además, la exclusión puede verse limitada por el propio desarrollo de la tecnología y los estándares de la industria para expresar adecuadamente la reserva de derechos.[40]

También debe tenerse en cuenta que, las obligaciones de transparencia no parecen ser claras o, al menos, responder a las demandas de la reserva de derechos. Es poco probable que la transparencia sea viable sin medidas tecnológicas eficaces y accesibles (Intellectual Property Office, Department For Science, Innovation & Technology And Department For Culture, Media & Sport, 2024). Como se ha explicado, el modelo de transparencia previsto en el Reglamento no incluye la identificación de obras particulares y sin el acceso real a un directorio de datos de formación con descripciones precisas la protección de las obras es imposible. Los titulares de los modelos de IA son reticentes a dar a conocer o compartir los datasets que utilizan para entrenarlos, inclusive, aun denominándose abiertos.[41] La transparencia algorítmica deviene en un eslabón fundamental en la materialización de la reserva de derechos y, en general, de la protección de los derechos de los autores en el entrenamiento de modelos de IA generativa.

Conclusiones

El Reglamento de IA no tiene como finalidad específica regular el tema de la IA generativa y los derechos de autor, sin embargo, establece herramientas jurídicas que podrían ser utilizadas para lograr la protección de estos derechos, como es la transparencia algorítmica. En el contexto de la IA generativa esta es parte esencial del sistema de protección de los derechos de autor y, concretamente de la materialización de la reserva de derechos, sin embargo, su configuración legal en el reglamento afecta su efectividad. La obligación de transparencia en relación con los datos de

36. Algunas de las medidas propuestas son: **1)** no rastrear sitios web que pongan a disposición contenidos que infrinjan los derechos de autor; **2)** respetar el Protocolo de Exclusión de Robots; **3)** identificar y cumplir con otras expresiones apropiadas de reservas de derechos, conforme los estándares de la industria, **4)** publicar información sobre el cumplimiento de la reserva de derechos, entre otras. (EU AI Office, 2024).
37. También conocido como raspado web, permite extraer datos de páginas web para su posterior almacenamiento y análisis. Es común que los rastreos web o *web scrapping* se identifiquen como parte de los datos públicos, así se reconoce en el *GPT-4o System Card*, se diferencian de los datos privados que no son más que aquellos procedentes de asociaciones de datos que no están disponibles públicamente, como contenidos de pago, archivos y metadatos. OpenAI (2024).
38. https://haveibeentrained.com/faq. Algunas empresas como OpenAI brindan la posibilidad de excluir voluntariamente el contenido para que no sea utilizado en el entrenamiento de los modelos, sin embargo, esta disposición hace referencia solo a los contenidos generados por la propia aplicación, función *Improve the model for everyone*). https://openai.com/policies/terms-of-use/.
39. Esto quedó demostrado en el caso Mumset de Reino Unido. Ante el contacto de la empresa con OpenAI para conceder una licencia por los contenidos, la respuesta fue que les interesaban los conjuntos de datos que no fueran accesibles en línea. (Mumsnet, (2024).
40. Así se propone en el Código de buenas prácticas de la IA para fines generales (EU AI Office, 2024). En el caso LAION, por ejemplo, se debatió si la exclusión en lenguaje natural establecida en la página web *HTML* era suficiente o debía existir una exclusión específica para los robots rastreadores que fuera suficientemente legible por las máquinas que recopilan datos con estos fines (Kneschke, 2024). El tribunal consideró que sí.
41. En los procesos judiciales que se llevan a cabo en el territorio estadounidense se han establecido un «Protocolo para el acceso a los datos de entrenamiento», con elevadas medidas de seguridad, sin acceso a Internet, a la red ni a otros ordenadores o dispositivos. (IN RE OPENAI CHATGPT LITIGATION, 2024).

entrenamiento no existe en relación con el uso de obras en particular, lo cual hace poco viable la protección y defensa del derecho de los autores en este contexto.

En el entrenamiento de los modelos de IA generales se materializa la contradicción entre el carácter territorial de los derechos de autor y la universalidad de la IA, lo cual determina la viabilidad y eficacia de las herramientas jurídicas previstas en el Reglamento de IA. La norma establece como regla general la obligación de los proveedores de modelos de IA de uso general de cumplir con la normatividad de derechos de autor de la Unión Europea, lo que incluye el cumplimiento de las obligaciones derivadas de la excepción del TDM. El régimen legal de TDM no es claro ni preciso desde un punto de vista técnico ni legal, lo que provoca dudas en relación con los posibles usos indebidos e infracciones de los derechos de autor, en particular el uso de obras como fuente del entrenamiento de los sistemas de IA generativos.

El sistema normativo de derechos de autor al que hace referencia el Reglamento de IA no ofrece soluciones jurídicas a los retos que esta tecnología de la IA generativa supone; por ello, es necesario adoptar normas de protección de los derechos que sean más claras con posterioridad a la entrada en vigor de la norma para lograr la consecución de los objetivos propuestos.

Referencias bibliográficas

COLE, S. (2024). «AI Video Generator Runway Trained on Thousands of YouTube Videos Without Permission». *404media* [en línea]. Disponible en: https://www.404media.co/runway-ai-image-generator-training-data-youtube/. [Fecha de consulta: 14 de agosto de 2024].

CRAWFORD, K., SCHULTZ, J. «Generative AI Is a Crisis for Copyright Law». *Issues in science and technology*. DOI: https://doi.org/10.58875/GUYG6120. [Fecha de consulta: 9 de agosto de 2024].

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. *Diario Oficial de las Comunidades Europeas*. L 77/20. 27. 3. 96 [en línea]. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:31996L0009. [Fecha de consulta: 9 de agosto de 2024].

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. *Diario Oficial de las Comunidades Europeas*. L 167/10. 22.6.2001 [en línea]. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32001L0029. [Fecha de consulta: 9 de agosto de 2024].

Directiva 2009/24/CE del Parlamento Europeo y del Consejo, de 23 de abril de 2009 , sobre la protección jurídica de programas de ordenador. *Diario Oficial de las Comunidades Europeas*. L 111/16. 5.5.2009. [en línea]. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32009L0024. [Fecha de consulta: 9 de agosto de 2024].

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. *Diario Oficial de la Unión Europea*. L 130/92, 17.05.2019. [en línea]. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=CELEX:32019L0790. [Fecha de consulta: 2 de agosto de 2024].

DORNIS, T. W.; STOBER, S. (2024). *Urheberrecht und Training generativer KI-Modelle*. Baden-Baden: Nomos. DOI: https://doi.org/10.5771/9783748949558-1. [Fecha de consulta: 22 de diciembre de 2024].

EUROPEAN INNOVATION COUNCIL y la SMES EXECUTIVE AGENCY (2024). «Artificial intelligence and copyright: use of generative AI tools to develop new content» [blog en línea]. Disponible en: https://intellectual-property-helpdesk.ec.europa.eu/news-events/news/artificial-intelligence-and-copyright-use-generative-ai-tools-develop-new-content-2024-07-16-0_en. [Fecha de consulta: 20 de julio de 2024].

EU AI Office (2024). «Second Draft of the General-Purpose AI Code of Practice» [en línea]. Disponible en: https://digital-strategy.ec.europa.eu/en/library/second-draft-general-purpose-ai-code-practice-published-written-independent-experts. [Fecha de consulta: 22 de diciembre de 2024].

GARCÍA VIDAL, Á. (2024). «Propiedad intelectual, minería de textos y datos y entrenamiento de la inteligencia artificial». *GA_P* [en línea]. Dispnible en: https://ga-p.com/wp-content/uploads/2024/10/Mineria_textos_datos.pdf. [Fecha de consulta: 22 de diciembre de 2024].

GONZÁLEZ OTERO, B. (2019). «Las excepciones de minería de textos y datos más allá de los derechos de autor: la ordenación privada contraataca». En: SÁIZ GARCÍA, C. y EVANGELIO LLORCA, R. *Propiedad Intelectual y Mercado Único Digital Europeo*. València: Tirant-Lo Blanch. DOI: https://doi.org/10.2139/ssrn.3477197

GUADAMUZ, A. (2023). «Photographer sues LAION for copyright infringement». *TechnoLlama* [en línea]. Disponible en: https://www.technollama.co.uk/photographer-sues-laion-for-copyright-infringement. [Fecha de consulta: 9 de agosto de 2024].

HAVE I BEEN TRAINED (s. f.). «Frequently Asked Questions». *Have I Been Trained?* [en línea]. Disponible en: https://haveibeentrained.com/faq. [Fecha de consulta: 9 de agosto de 2024].

OPENAI (s. f.). «Europe Terms of Use». *OpenAI* [en línea]. Disponible en: https://openai.com/policies/terms-of-use/. [Fecha de consulta: 15 de agosto de 2024].

HUTIRI, W.; PAPAKYRIAKOPOULOS, O.; XIANG, A. (2024). «Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators». En: *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. New York: Association for Computing Machinery, págs. 359-376. DOI: https://doi.org/10.1145/3630106.3658911. [Fecha de consulta: 2 de agosto de 2024].

IN RE OPENAI CHATGPT LITIGATION (2024). Training Data Inspection Protocol. United States District Court. Northern District of California San Francisco Division. Master File Case No. 3:23-CV-03223, [en línea]. Disponible en: https://app.ediscoveryassistant.com/case_law/59943-in-re-openai-chatgpt-litig. [Fecha de consulta: 2 de diciembre de 2024].

INTELLECTUAL PROPERTY OFFICE, DEPARTMENT FOR SCIENCE, INNOVATION & TECHNOLOGY AND DEPARTMENT FOR CULTURE, MEDIA & SPORT (2024). «Open consultation. Copyright and Artificial Intelligence». *Gov.uk* [en línea]. Disponible en: https://www.gov.uk/government/consultations/copyright-and-artificial-intelligence/copyright-and-artificial-intelligence. [Fecha de consulta: 22 de diciembre de 2024].

JIMÉNEZ SERRANÍA, V. (2024). «Medidas de apoyo a la innovación y arquitectura de gobernanza». En: JIMÉNEZ SERRANÍA, V., CASTILLA BAREA, M., MÍGUEZ MACHO, L, BARRIO ANDRÉS, M., DELGADO MARTÍN, J., MUÑOZ GARCÍA, C., & TORRES CARLOS, M. (2024). *El Reglamento Europeo de Inteligencia Artificial*. València: Tirant lo Blanch, págs. 111-138 [en línea]. Disponible en: https://biblioteca-nubedelectura-com.eu1.proxy.openathens.net/cloudLibrary/ebook/info/9788410713048. [Fecha de consulta: 20 de julio de 2024].

JONES, E. (2023). «What is a foundation model?». *Ada Lovelace Institute* [en línea]. Disponible en: https://www.adalovelaceinstitute.org/resource/foundation-models-explainer/. [Fecha de consulta: 20 de junio de 2024].

KNESCHKE, R. (2024). «Was ist erlaubt beim Erstellen von KI-Trainingsdaten? Erster Verhandlungstag im Verfahren gegen LAION e.V.» Traducción realizada con la versión gratuita del traductor DeepL.com. *Alltag eines Fotoproduzenten* [en línea]. Disponible en: https://www.alltageinesfotoproduzenten.de/2024/07/12/was-ist-erlaubt-beim-erstellen-von-ki-trainingsdaten-erster-verhandlungstag-im-verfahren-gegen-laion-e-v/. [Fecha de consulta: 15 de agosto de 2024].

MEZEI, P. (2024). «A Saviour or a dead end? Reservation of rights in the age of generative AI». *European Intellectual Property Review*, vol. 46, n.º 7, págs. 461-469. DOI: https://doi.org/10.2139/ssrn.4695119. [Fecha de consulta: 9 de agosto de 2024].

MUMSNET, J. (2024). «Why we're taking legal action against Open AI and other scrapers». *Mumsnet* [en línea]. Disponible en: https://www.mumsnet.com/talk/site_stuff/5122770-why-were-taking-legal-action-against-open-ai-and-other-scrapers. [Fecha de consulta: 15 de agosto de 2024].

OPENAI (2023). «Written evidence (LLM0113) House of Lords Communications and Digital Select Committee inquiry: Large language models». *Parliament.uk* [en línea]. Disponible en: https://committees.parliament.uk/writtenevidence/126981/pdf/. [Fecha de consulta: 20 de febrero de 2024].

OPENAI (2024). «GPT-4o System Card». *OpenAI* [en línea]. Disponible en: https://cdn.openai.com/gpt-4o-system-card.pdf. [Fecha de consulta: 9 de agosto de 2024].

OCDE (2024). «Recommendation of the Council on Artificial Intelligence». OECD/LEGAL/0463 [en línea]. Disponible en: https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449. [Fecha de consulta: 9 de agosto de 2024].

Resolución del Parlamento Europeo, de 20 de octubre de 2020, sobre los derechos de propiedad intelectual para el desarrollo de las tecnologías relativas a la inteligencia artificial (2020/2015(INI) [en línea]. Disponible en: https://www.europarl.europa.eu/doceo/document/TA-9-2020-0277_ES.html. [Fecha de consulta: 2 de agosto de 2024].

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. *Diario Oficial de la Unión Europea*, 12.07.2024 [en línea]. Disponible en: https://eur-lex.europa.eu/legal-content/ES/TXT/PDF/?uri=OJ:L_202401689. [Fecha de consulta: 2 de agosto de 2024].

SAG, M. (2019). «The New Legal Landscape for Text Mining and Machine Learning». *Journal of the Copyright Society of the USA*, vol. 66, pág. 291 [en línea]. Disponible en: https://ssrn.com/abstract=3331606. [Fecha de consulta: 23 de noviembre de 2023].

SHEN, M. (2024). «Rethinking Data Selection for Supervised Fine-Tuning». *arXiv* [en línea]. Disponible en: https://arxiv.org/pdf/2402.06094. [Fecha de consulta: 14 de agosto de 2024].

SHIKSHA ONLINE (2023). «Text Mining in Data Mining». *Shiksha online* [en línea]. Disponible en: https://www.shiksha.com/online-courses/articles/text-mining-in-data-mining/. [Fecha de consulta: 23 de julio de 2023].

Sentencia 310 o 227/23. Tribunal Regional de Hamburgo, Sala de lo Civil 10, 27 de septiembre de 2024.

UNESCO (2021). Recomendación sobre la Ética de la Inteligencia Artificial, Paris [en línea]. Disponible en: https://www.unesco.org/es/legal-affairs/recommendation-ethics-artificial-intelligence. [Fecha de consulta: 2 de agosto de 2024].

UMG Recordings, Inc. v. Uncharted Labs, Inc. (1:24-cv-04777), District Court, S.D. New York (2024), [en línea]. Disponible en: https://storage.courtlistener.com/recap/gov.uscourts.nysd.623701/gov.uscourts.nysd.623701.26.0.pdf. [Fecha de consulta: 14 de agosto de 2024].

Cita recomendada

ORDELIN FONT, Jorge Luis (2025). «Derechos de autor y entrenamiento de sistemas de IA generativos: las obligaciones de transparencia y la minería de textos y datos en la normativa europea». *IDP. Revista de Internet, Derecho y Política*, núm. 42. UOC. [Fecha de consulta: dd/mm/aa]. DOI: http://dx.doi.org/10.7238/idp.v0i42.431327

Sobre la autoría

Jorge Luis Ordelin Font
Centro de Investigación y Docencia Económicas (CIDE), México
jorge.ordelin@cide.edu

Profesor investigador titular de Propiedad Intelectual y Nuevas Tecnologías de la División de Estudios Jurídicos del Centro de Investigación y Docencia Económicas (CIDE), México. Investigador nacional nivel I del Sistema Nacional de Investigadores, México. Profesor de la maestría en Derecho y TIC del Centro de Investigación e Innovación en TIC (INFOTEC), México. Conferencista invitado de la Organización Mundial de la Propiedad Intelectual (OMPI) en los ámbitos del derecho de autor y nuevas tecnologías. Miembro de la Línea de Investigación de Derecho e Inteligencia Artificial del Instituto de Investigaciones Jurídicas de la UNAM y experto de la cátedra Iberoamericana de Cultura Digital y Propiedad Intelectual, promovida por la Organización de Estados Iberoamericanos para la Educación, la Ciencia y la Cultura, en colaboración con la Universidad de Alicante. Consultor en temas de propiedad intelectual, inteligencia artificial y nuevas tecnologías.