

REPÚBLICA BOLIVARIANA DE VENEZUELA
UNIVERSIDAD MONTEÁVILA
COMITÉ DE ESTUDIOS DE POSTGRADO
ESPECIALIZACIÓN EN PROPIEDAD INTELECTUAL

**El entrenamiento de modelos de inteligencia artificial generativa con obras protegidas
por copyright: límites del fair use en el derecho estadounidense.**

Abg. Omar Alfredo Martinez Guere, C.I 27.343.197

Tutor: **Mag.** William Olivero

Caracas, 27 de mayo de 2026.

Resumen.....	1
Abstract:.....	2
INTRODUCCIÓN.....	4
CAPÍTULO I.....	6
EL PROBLEMA.....	6
Planteamiento del problema.....	7
Objetivos.....	8
Objetivo General:.....	8
Objetivos Específicos:.....	8
Justificación y delimitación.....	9
CAPÍTULO II.....	11
MARCO TEÓRICO.....	11
Antecedentes.....	11
Bases teóricas.....	12
Copyright.....	12
Inteligencia artificial.....	17
Machine Learning, Deep Learning, Redes Neuronales Artificiales (RNA), Foundation Models, Large Language Models.....	19
Crawlers, Common Crawl, Books3 y shadows libraries en la recolección de datos por empresas de Inteligencia Artificial Generativa.....	21
Copyright, machine learning y minería de texto y datos.....	30
Bases legales.....	35
Copyright Act.....	35
Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson v. Anthropic pbc.....	36
New York Times: noticias en el alimento de inteligencia artificial generativa.....	48
Kadrey et al. v. Meta Platforms.....	50
CAPÍTULO III: MARCO METODOLÓGICO.....	51
Línea de investigación.....	52
Técnicas y herramientas de recolección y procesamiento de datos.....	53
Presentación y análisis de resultados.....	54
CAPÍTULO IV: DESARROLLO DE LA PROPUESTA.....	55
CAPÍTULO V.....	60
RECOMENDACIONES Y CONCLUSIONES.....	60
REFERENCIAS.....	62

Resumen

El avance tecnológico ha implicado una carrera veloz entre las grandes empresas de Estados Unidos con el fin de poder establecerse con éxito en el mercado de las inteligencias artificiales generativas. Esta competencia se diluye en: ¿cómo entrenar estas tecnologías - también llamadas Large Language Models (LLMs)- para tener una ventaja competitiva?. El presente trabajo consiste en analizar aquellos casos estadounidense donde empresas tecnológicas hayan tomado obras protegidas por copyright para alimentar sus sistemas, con el fin de para poder garantizar a sus usuarios una mejor experiencia.

Actualmente la base de datos de Edward Lee del año 2026, menciona un total de 106 casos activos en materia de *copyright* y empresas de inteligencia artificial. Al no tener hasta el momento de este trabajo una decisión firme que sirva como estándares mínimos, surge la urgencia de seguir hablando del tema y aportar a esta línea de trabajo que se encuentra envuelta en disyuntivas, para poder estar preparados antes más sucesos de esta índole, permitiendo tanto el proceso tecnológico como el respeto a los titulares de derechos.

Palabras claves: Inteligencia Artificial, Gen IA, Machine Learning, Deep learning, Crawlers, Copyright, fair use, Large Language Models, Foundation Models

Abstract:

Recent technological advances have triggered a race among major U.S. technology companies to secure a leading position in the rapidly growing generative artificial intelligence market. The key question in this matter is: how can these technologies—commonly known as Large Language Models (LLMs)—be trained in a way that provides a competitive advantage?

This study examines U.S. copyright cases involving technology companies that allegedly used copyrighted works to train their artificial intelligence systems. By relying on large amounts of existing data, these companies seek to improve the capabilities, accuracy, and overall performance of their models, thereby enhancing the user experience.

According to 2026 Edward Lee's database, there are currently 106 active copyright lawsuits involving artificial intelligence companies. Because no final judicial decision has yet established clear legal standards governing the use of copyrighted works for AI training, the issue remains highly contested. This creates a need for continued academic and legal discussion. By analyzing these cases, this research aims to contribute to a developing area of law marked by significant uncertainty and competing interests. Ultimately, the challenge is to find an appropriate balance between encouraging technological innovation and safeguarding the rights of copyright holders.

Keywords: Artificial Intelligence, Gen AI, Machine Learning, Deep Learning, Crawlers, Copyright, Fair Use, Large Language Models, Foundation Models

Abreviaturas

LLMs: Large Language Models

GenAI: Generative Artificial Intelligence

AI: Artificial Intelligence

ML: Machine Learning

FM: Foundation Models

DL: Deep Learning

RNA: Redes Neuronales Artificiales

TDM: Text and Data Mining

RLHF: Reinforcement Learning from Human Feedba

INTRODUCCIÓN

"Are we losing ourselves and our own essence?"

Are we being overwhelmed by technology?"

H.R Giger

Si nos montamos en el famoso DeLorean DMC-12 y viajamos al siglo XVIII, podremos presenciar el enorme contraste que se ofrecía entonces en la industria literaria. En aquellos años:

"Se probaba un libro al igual que se podría haber catado un vaso de vino, pues el lector observaba la impresión sobre el papel, no sólo por medio de su significado. Y cada vez que había poseído el libro en todas sus características físicas, se sentaba a leerlo" (Darnton, 2003, p. 254).

Hoy, en cambio, el lector se sienta frente a una computadora o un dispositivo móvil, selecciona el sistema de Inteligencia Artificial (IA) de su preferencia, estira los dedos sobre el teclado y escribe un *prompt* para que ese sistema le resuma el libro que desea conocer. En el camino, queda relegada la otredad de quien quiso dejar una parte de sí en la obra, ya que, en cada texto escrito subsiste la esencia del autor.

En la actualidad las empresas de inteligencia artificial generativas (GenAI) ocupan una gran parte de la cotidianidad de la sociedad moderna, se han convertido en los nuevos educadores capaces de resumir libro complejos y completos con un solo *prompt*, el cual

suministra toda la información detallada de lo solicitado de forma concisa que satisface al usuario, sin preguntarnos de dónde viene lo que estamos leyendo.

En esta escena cotidiana de escribir un *prompt* se esconde una cuestión jurídica relevante: la forma en que estos sistemas de GenAI se alimentan —es decir, el modo en que las empresas recaudan la información necesaria para entrenar dichas tecnologías—, suele pasar desapercibida para el usuario común. Resulta indispensable, indagar sobre cómo han alimentado estos sistemas durante los últimos años, y si ese conjunto de conocimientos ha sido obtenido en detrimento del *copyright* o si, por el contrario, encuentra refugio en la figura del *fair use* propia del sistema estadounidense.

CAPÍTULO I.

EL PROBLEMA

La revolución de la IA se nos presenta de forma utópica y de absoluta necesidad para el +ser humano, sin embargo “ninguna utopía es completamente real o completamente razonable. Existe una manzana prohibida en cada jardín, un gusano en cada manzana”. (Shaviro, 2014, p-55). La utopía que se construye en torno a la GenAI, sostenida en parte por las estrategias de marketing del sector tecnológico, no se corresponde con la realidad que estas ofrecen.

Actualmente, genera honda preocupación el método mediante el cual las empresas tecnológicas nutren sus Large Language Models (LLMs) para comercializar sus GenAI. El desarrollo y optimización de estos sistemas depende, en parte, directamente de la producción creativa de miles de autores, cuyas obras han sido indexadas sin autorización ni retribución

económica a través de técnicas de recolección masiva de datos como el *crawling* o *shadow libraries*. Como consecuencia, la sociedad contemporánea se enfrenta a un escenario donde los frutos de la creatividad humana corren el riesgo de ser capitalizados por corporaciones estadounidenses tecnológicas en detrimento de los derechos de sus creadores.

De esta problemática se desprende el objeto de estudio del presente Trabajo Especial de Grado (TEG), cuyo propósito es contribuir al debate actual sobre este fenómeno y proponer una interpretación razonable respecto al rumbo que debe tomar la respuesta jurídica; todo ello sin perder de vista el elemento que constituye la esencia de la condición humana: la creatividad.

Planteamiento del problema

En consecuencia de los sistemas de GenAI, actualmente nos encontramos en un punto de inflexión que amerita de peculiar análisis; como comenta Fenollosa:

“Las grandes revoluciones se producen en dos fases. Primero, gradualmente. Pequeños cambios casi imperceptibles, avances incrementales que solo interesan a los expertos y pasan por el mundo sin generar titulares. Y entonces, de repente, el punto de inflexión, algunas veces causado por un detonante mientras que otras es sencillamente la pequeña gota que colma el vaso” (Fenollosa, 2024, p-11);

Hoy estamos frente a una tecnología nunca antes vista, la cual ha logrado integrarse a la sociedad de forma acelerada y con poca regulación. Es por esto que el aprendizaje de las máquinas mediante *machine learning* (ML) genera hoy en día una serie de desafíos a tener en

cuenta, ya que, muchas de los datos e información usadas para entrenar a las GenAI por parte de estas grandes empresas de IA, se encuentran protegidas por el *copyright*.

Un claro ejemplo de esta problemática lo podemos ver en la IA titulada “*The Next Rembrandt*”, la cual fue entrenada mediante ML con una cantidad elevada de obras del pintor holandés Rembrandt, logrando así, crear por su cuenta un nuevo cuadro, resultado de todos los datos que le fueron suministrados. En este caso, no se encuentra mayor conflicto con el *copyright*, ya que, el pintor corresponde al siglo XVII, y dichas obras ya pertenecen al dominio público, es por eso que se considera que en su entrenamiento no se vulnera ningún derecho.

Sin embargo, ¿qué sucede si tomamos el mismo procedimiento en el cual la IA es entrenada con datos y obras protegidas por el *copyright*? ¿Estamos aquí frente a un *fair use*? O en efecto, ¿estamos ante una vulneración de derechos y un abuso por parte de las empresas de AI?

Es por esto que surge la necesidad de traer este tema al debate, y poder delimitar hasta dónde puede llegar el entrenamiento de las GenAI y cuáles pueden ser las alternativas a tomar en cuenta por parte de las empresas de IA o los propietarios de derechos.

Objetivos

Objetivo General:

Analizar las implicaciones legales del entrenamiento de modelos de inteligencia artificial generativa mediante el uso de obras protegidas por *copyright* en el derecho estadounidense, a

fin de identificar los criterios que delimitan la aplicación del fair use y sus efectos en el derecho venezolano.

Objetivos Específicos:

1.- Analizar el marco legal de los Estados Unidos en materia de *copyright*, con el fin de identificar criterios jurídicos que permitan delimitar el *crawling* y las *shadow libraries*.

2.- Revisar precedentes judiciales en el sistema de Estados Unidos para establecer los fundamentos legales utilizados para calificar el *fair use* por parte de las empresas de inteligencia artificial generativas.

3.- Determinar las posibles consecuencias del *crawling* y *shadow libraries* en los titulares de los derechos en el sistema de los Estados Unidos.

Justificación y delimitación

La presente investigación tiene sus raíces en la acelerada expansión de las GenAI y en el controvertido proceso de entrenamiento de estos sistemas mediante técnicas de *crawling* y el uso de *shadow libraries* sobre obras protegidas por *copyright*. En la actualidad, se observa una progresiva difuminación de atributos tradicionalmente vinculados a la creación humana, tales como la creatividad y el ingenio; como consecuencia de la manera en que las Redes Neuronales Artificiales (RNA) procesan, estas absorben grandes volúmenes de contenido protegido sin el consentimiento expreso ni la correspondiente remuneración a los titulares de derechos.

Desde un primer análisis se puede considerar la figura del *fair use* en la alimentación de los sistemas de GenAI; esta figura legal consiste en la “doctrina que permite promover la libertad de expresión permitiendo el uso de material sin licencia de obras protegidas por el sistema de *copyright* en ciertas circunstancias”(.S. Copyright Office Fair Use Index. S/f). Sin embargo, las grandes empresas de tecnología hacen uso de esta figura para poder evadir la remuneración a los titulares de los derechos para usar sus obras, a pesar de que no siempre se pueda justificar sus acciones.

Por lo tanto, resulta preciso examinar las sentencias del sistema de EE.UU que están constituyendo pilares fundamentales para la configuración de la doctrina del fair use. Asimismo, es imperativo abordar litigios recientes como Bartz v. Anthropic, los cuales ofrecen una primera perspectiva sobre la manera en que los tribunales norteamericanos han comenzado a articular la relación entre el avance tecnológico y la protección otorgada por el sistema de copyright.

Dilucidar si un sistema de GenAI posee inteligencia real, o conciencia resulta estéril para los fines de esta investigación, pues ello implicaría desviar la atención del problema jurídico central. Por el contrario, la presente propuesta busca reposicionar a la persona en el centro del debate sobre IA, examinando cómo el despliegue de estas tecnologías puede impactar negativamente tanto en la sociedad como en atributos intrínsecamente humanos, entre ellos: la creatividad y el ingenio. El propósito de este TEG no consiste en debatir si plataformas como ChatGPT u otros modelos de GenAI son creativos o merecedores de protección jurídica, sino en analizar el progresivo desplazamiento del valor económico y cultural de las creaciones humanas hacia los activos comerciales de las empresas tecnológicas.

No constituyen objeto de estudio del presente TEG los avances científicos y tecnológicos impulsados mediante el uso de crawlers y técnicas de minería de textos y datos (TDM). Por el contrario, el enfoque de esta investigación se dirige exclusivamente hacia las empresas desarrolladoras de GenIA, cuyos sistemas permiten, a partir de una simple instrucción o *prompt*, procesar, sintetizar e incluso resumir obras literarias completas protegidas por *copyright*.

CAPÍTULO II.

MARCO TEÓRICO

Antecedentes

A los fines de la presente investigación, el impacto de las GenAI reviste especial relevancia, por cuanto la lógica de inmediatez propia del entorno digital contemporáneo ha encontrado una de sus máximas expresiones en estas plataformas tecnológicas. En efecto, dichos sistemas permiten que, mediante una simple instrucción o *prompt*, pueda automatizarse la síntesis del contenido esencial de una obra literaria, desplazando progresivamente la necesidad de invertir tiempo en su lectura o recursos económicos en la adquisición del ejemplar.

Esta problemática plantea interrogantes fundamentales: ¿cuál es el origen de los datos que nutren a las GenAI?, ¿se están entrenando estos grandes LLMs mediante el uso de material protegido por *copyright*?, y ¿pueden tales conductas encuadrarse dentro de la doctrina del *fair use*?. La ausencia de consenso en torno a estas cuestiones resulta evidente en litigios de gran relevancia como *Bartz v. Anthropic*, *The New York Times Co. v. Microsoft Corp. & OpenAI* y *Kadrey v. Meta Platforms, Inc.*, casos emblemáticos que serán objeto de análisis en la presente investigación.

Bases teóricas

Copyright

Debido a la influencia jurídica de Inglaterra, el sistema de EE.UU se adscribe al modelo del *common law*, el cual se rige por la doctrina del *copyright*. Históricamente, esta institución fue concebida, en sentido estricto, como el derecho exclusivo de controlar la reproducción de las obras protegidas. Este entramado normativo tiene sus orígenes en el Estatuto de la Reina Ana de 1710, instrumento que se centraba “en la copia y no en el autor, con las múltiples consecuencias que de ello se derivan” (Miró, 2007, p-125). Bajo esta concepción originaria, el autor era entendido como un individuo autónomo cuya creación intelectual daba lugar al nacimiento de un derecho de propiedad sobre la obra producida.

En base a esta última premisa, el *copyright* se funda “en el concepto del individuo único que crea algo original y tiene derecho a obtener ganancias de ese trabajo” (Mark, 1993, p-2). En los albores de este sistema se usaba la analogía del autor con el concepto de paternidad, “espejo de la sociedad patriarcal que se preocupaba por el linaje, la dinastía y la sangre”.(Mark, 1993, p-39). De este concepto surge la palabra *plagio* que proviene del latín *plagiare* que significa secuestro (Diccionario etimológico de Chile); he aquí las connotaciones que se le daba al autor de una obra literaria y sus primeros pasos como propiedad. Mark Rose en su libro “Authors and Owners”, nos asoma una cita de cómo era asimilados estos conceptos en 1710:

“Un libro es propiedad del autor, es el hijo de su invención, el fruto de su cerebro; si vende su propiedad, ésta pasa a ser derecho del comprador; si no, es tan suyo como suyos son su esposa y sus hijos. Pero he aquí, en esta nación cristiana, estos hijos de nuestras cabezas son apresados, cautivados, secuestrados y llevados en cautiverio, y no hay nadie que los redima.”
(Mark, 1993, p-39)

Una vez dicho esto es de suma importancia definir la originalidad, ya que, es el pilar que sostiene todo el sistema anglosajón del *copyright*, esta se puede entender como cualquier obra que plasme la esencia del autor, una creación que incorpora un grado mínimo de creatividad y se diferencia de las expresiones previamente existentes; la sección 102 del *Copyright Act* establece:

“La protección del copyright subsiste, de conformidad con este título, en obras originales con autoría fijadas en cualquier medio tangible de expresión, ahora conocido o desarrollados posteriormente, a partir del cual puedan ser percibidos, reproducidos o de otro modo comunicado, ya sea directamente o con la ayuda de una máquina o dispositivo. Bajo ningún término la protección se extiende a las ideas usadas en la elaboración del trabajo”. (Copyright Act. Sec 102).

Sin embargo, en el sistema estadounidense el concepto de originalidad no posee un carácter absoluto, sino que se atribuye a toda aquella obra que incorpore un mínimo de creatividad humana. Este estándar quedó firmemente consolidado en el célebre precedente de *Feist Publications, Inc. v. Rural Telephone Service Co.*, en el cual la Corte Suprema de los Estados Unidos delimitó los requisitos mínimos necesarios para otorgar protección jurídica a una obra, estableciendo:

“Para que una obra califique en la protección del copyright, esta debe ser original del autor. Sin duda, el nivel requerido de creatividad es extremadamente bajo; incluso una pequeña cantidad será suficiente. La gran mayoría de las obras superan la calificación con bastante facilidad, ya que poseen cierta chispa creativa, por muy cruda, u obvia que pueda ser. Originalidad no significa novedad; una obra puede ser original aunque se parezca mucho a otras obras, siempre que la similitud sea fortuita y no resultado de una copia” (Feist Publications, Inc. v. Rural Telephone Service Co., 1991).

Al tenor de lo expresado por dicho Tribunal, la originalidad se materializa en la manera en que el autor plasma su impronta personal en un medio tangible. Bajo esta premisa, es posible encontrar una multiplicidad de poemas que abordan temas como el amor o la melancolía —desde José Asunción Silva hasta Novalis—, donde cada obra posee ese grado mínimo de creatividad que le otorga singularidad y la diferencia de las demás, descartando así cualquier supuesto de plagio.

Adicionalmente, Mark Rose nos ofrece la siguiente cita que nos lleva a como se consideraba la originalidad en los inicios del copyright durante el siglo XVIII:

“No hay dos caras ni dos mentes iguales; pero todos llevan la evidente marca de separación de la Naturaleza. Nuestro derecho de nacimiento es la distinción, y sin embargo esta distinción no dura: Nacidos Originales, ¿cómo es posible que morimos siendo Copias?. Desafortunadamente, perdemos nuestra originalidad nativa cuando esa imitación entrometida arrebató la pluma y borra la marca de separación de la naturaleza, cancela su amable intención, destruye toda Individualidad mental. La naturaleza, entonces, es una especie de autores y nosotros somos sus obras: cada uno de nosotros es un texto original producido por un genio creador; la marca visible de originalidad es el rostro humano; y esto a su vez es el signo de la distinción invisible de la mente. Como un texto original, la persona producida por la naturaleza está destinada a ser un autor original” (Mark, 1993, p-120).

Con base a lo anterior, se evidencia cómo, a lo largo de la evolución del *copyright*, se ha procurado proteger las obras del ingenio humano con el propósito de preservar la identidad del creador y fomentar el desarrollo del mercado cultural a través del requisito de la originalidad. Esta última noción se concibe, en términos de Hariani (2011), bajo la premisa de que es original “todo aquello que sea esfuerzo del propio autor” (p-8). De este modo, se entiende que el sistema del *copyright* se fundamenta en principios orientados a salvaguardar la individualidad de cada sujeto.

Fair use

En síntesis, el *copyright* surge ante la necesidad de reconocer a los autores en un entorno de progresiva globalización, donde el comercio de obras intelectuales comenzaba a expandirse de manera acelerada. El propósito del *common law* respecto a los bienes intangibles no consiste en la creación de monopolios que impidan el libre acceso a la información; por el contrario, este sistema busca delimitar la frontera entre lo privado y lo público, reconociendo la originalidad de las obras al tiempo que promueve el desarrollo de la economía de la creatividad.

Es por esto que el Copyright Act en su sección 107 establece 4 principios que rigen la figura del *fair use*:

“(1) El propósito y carácter del uso, incluido si dicho uso es de naturaleza comercial o tiene fines educativos sin fines de lucro; (2) la naturaleza del trabajo protegido por copyright ; (3) la cantidad y sustancialidad de la parte utilizada en relación con la obra protegida por derechos de autor en su conjunto; y (4) el efecto del uso sobre el mercado potencial o el valor de la obra protegida por copyright. El hecho de que una obra sea inédita no impedirá en sí misma una conclusión de uso legítimo si tal hallazgo se hace tras considerar todos los factores anteriores”
(17 U.S.C. § 107)

Lo que se puede interpretar de la ley citada, sería, que “el Acta define la naturaleza del uso específico como el propósito y carácter del uso, incluyendo si tal uso es de naturaleza comercial o es para educación sin fines de lucro”(Patterson, 1997, p-43)

Dicho esto, el *fair use* se basa en el uso legítimo que se le pueda dar a una obra protegida por el *copyright*, las bases legales que sustentan esta normativa serán analizadas a fondo más adelante.

Inteligencia artificial

Antes de analizar a fondo el impacto jurídico de las GenAI, conviene indagar un poco sobre sus fundamentos para entender su impacto actual, por lo tanto, Fenollosa comenta el término global de IA como:

“Una disciplina cuyo objetivo es crear inteligencia sintética, artificial, con el objetivo de que una máquina se comporte de manera inteligente. Hoy día se entiende popularmente como la tecnología que consigue que los ordenadores realicen tareas complejas, a veces como persona, a veces de manera distinta, mejor o peor que nosotros”(Fenollosa, 2024, p-27).

La IA no es realmente un concepto de este siglo, sin embargo, en estos últimos años ha llegado a un punto donde sus avances nos resultan sorprendentes e indistinguibles a novelas de ciencia ficción. La primera vez que se usó el término moderno IA fue en 1950 por el británico Alan Mathison Turing; este fue el primero en hipotetizar que “el cerebro humano es en gran medida una computadora. Teorizó que el córtex al nacimiento es una máquina desorganizada, que con entrenamiento puede convertirse en organizada” (Copeland, 2026). Además se dio nacimiento a la Prueba de Turing, esta se basa en:

“Un ejercicio muy simple: dejar que un humano mantenga una conversación a ciegas con dos interlocutores desconocidos, por ejemplo a través de un chat de ordenador o de un intermediario. Si el interrogador es incapaz de determinar quién es el humano y quién es la máquina, entonces esa aplicación ha superado la prueba y merece la etiqueta de IA” (Fenollosa, 2024, p-31).

Actualmente existen variadas formas de explicar a la IA, Lasse Rouhiainen, se aproxima este concepto como:

“La habilidad de los ordenadores para hacer actividades que normalmente requieren inteligencia humana.[...] la IA es la capacidad de las máquinas para usar algoritmos, aprender de los datos y utilizar lo aprendido en la toma de decisiones tal y como lo haría un ser humano. Sin embargo, a diferencia de las personas, los dispositivos basados en IA no necesitan descansar y pueden analizar grandes volúmenes de información a la vez. Asimismo, la proporción de errores es significativamente menor en las máquinas”. (Rouhiainen, 2018, p-17).

En EE.UU coexisten diversos criterios para conceptualizar la IA; no obstante, una de las aproximaciones más precisas es la propuesta por el *National Science and Technology Council* (2019), el cual la define como aquella tecnología que “permite a las computadoras y a otros sistemas automatizados realizar tareas que históricamente han requerido de la cognición humana, así como de facultades que típicamente consideramos habilidades humanas para la toma de decisiones” (National Science and Technology Council, 2019, p-1).

Machine Learning, Deep Learning, Redes Neuronales Artificiales (RNA), Foundation Models, Large Language Models.

Todo esto nos lleva a tratar de aproximarnos al concepto de ML, el cual “se trata de un aspecto de la informática en el que los ordenadores o las máquinas tienen la capacidad de aprender sin estar programados para ello. Un resultado típico serían las sugerencias o predicciones en una situación particular” (Rouhiainen, 2018, p-19).

Bajo una estructura análoga a la de una muñeca matrioska, dentro de este proceso se ubica el aprendizaje profundo o *deep learning* (DL). Este modelo consiste en la capacidad de ejecutar tareas complejas de manera automatizada gracias al procesamiento de volúmenes masivos de datos y a la implementación de RNA. Estas redes neuronales son sistemas computacionales y matemáticos diseñados para simular la estructura y el funcionamiento biológico de las neuronas humanas, Matich las define como:

“Compuestos por un gran número de elementos de procesos muy interconectados, los cuales procesan información por medio de su estado dinámico como respuesta a entradas externas, también se podría considerar como redes interconectadas masivamente en paralelo de elementos simples (usualmente adaptativos) y con organización jerárquica, las cuales intentan interactuar con los objetos del mundo real del mismo modo que lo hace el sistema nervioso biológico” (Matich, 2001, p-8).

Ahora bien, es necesario mencionar un brevemente sobre los Foundation Models (FM), estos son “un tipo de modelo de IA que se pre entrena con una gran cantidad de datos para realizar una serie de tareas”(Google Cloud, s. f.), estos son entrenados tanto con imágenes, videos, canciones, texto, entre otros datos. Es de estos modelos que surgen los LLMs los cuales son:

“Una categoría de modelos de deep learning entrenados sobre inmensas cantidades de datos, lo que los hace capaces de comprender y generar lenguaje natural y otros tipos de contenidos para realizar una amplia gama de tareas. Los LLMs se basan en un tipo de arquitectura de red neuronal llamada transformador que destaca en la gestión de secuencias de palabras y la captura de patrones en el texto” (Stryker. s.f).

Para entender el concepto de FM y LLMs, es importante saber que “todos los LLMs son FM pero no todos los FM son LLMs” (Google Cloud, s. f.).

En consecuencia, lo que actualmente vemos en Chat GPT, Deep Seek, Gemini, Copilot, Claude, entre otros; es conocido como GenAI. El punto de inflexión para la creación de estos sistemas se debe gracias a la empresa Google y la implementación del transformador en los LLMs, incluso el nombre de Chat GPT proviene de esta tecnología: *Generative Pre-trained Transformer*.

“En 2017 el proyecto de Google, fue bautizado con un nombre suficientemente descriptivo: transforma los datos. El transformador es la siguiente iteración del aprendizaje profundo (Deep learning), un tipo de red neuronal artificial pensada sobre todo para traducir textos. Los transformadores no traducen, sino que interpretan. Es como un humano bilingüe: no convierte las frases de un idioma a otro, sino que ha creado su propio lenguaje simbólico interno, su esperanto, en el que almacena ideas y en el que piensa nativamente” (Fenollosa, 2024, p-73).

En términos generales, la IA puede entenderse como el concepto matriz cuyas bases teóricas comenzaron a consolidarse a partir de los aportes realizados por Alan Turing en 1950. Desde entonces, se desarrolló una evolución progresiva de tecnologías y modelos interrelacionados que abarcan el ML, el DL, los FM, los LLMs y, más recientemente, la GenAI. De esta manera, cada etapa tecnológica se construye sobre la anterior, optimizando sus capacidades y ampliando progresivamente el alcance de éstas.

Crawlers, Common Crawl, Books3 y shadows libraries en la recolección de datos por empresas de Inteligencia Artificial Generativa.

Es indispensable abordar el funcionamiento de los *crawlers* y de plataformas públicas como *Common Crawl* para comprender adecuadamente los mecanismos de recopilación de datos utilizados en el entrenamiento de sistemas de GenAI; estos se entienden como:

“Los Crawlers son programas automatizados que visitan sitios web y descargan su contenido. En la era de la IA, las empresas utilizan rastreadores para diversos fines. En el momento de escribir este artículo, existen tres tipos principales de rastreadores relacionados con la IA: (1) rastreadores para recopilar datos de entrenamiento (por ejemplo, GPTBot de OpenAI), (2) rastreadores para aumentar los asistentes respaldados por IA (por ejemplo, ChatGPT-User de OpenAI) y (3) rastreadores para facilitar motores de búsqueda respaldados por IA, por ejemplo, SearchBot de OpenAI” (Enze Liu et al, 2025, p-79).

Con fines del presente TEG, nos enfocaremos en el primer punto: *AI data crawlers*. “Un uso importante de los rastreadores es recopilar datos para entrenar modelos de IA. Algunas empresas han desarrollado sus propios rastreadores para tales fines y otras dependen de rastreadores de terceros, por ejemplo, *Common Crawl* (Enze Liu et al, 2025, p-79).

Las palabras de Meredith Whittaker, presidenta del *Signal Technology Foundation*, resultan de relevancia:

“Los actuales modelos de IA abiertos suelen depender de dos repositorios de datos extraídos de internet y puestos a disposición para su uso generalizado: el conjunto de datos Common Crawl y The Pile. Common Crawl es un repositorio de datos extraídos de internet que una organización homónima y sin ánimo de lucro, con sede en Estados Unidos, actualiza cada mes con el objetivo de democratizar [...] los datos para que todo el mundo, no solo las grandes empresas, pueda desarrollar investigación y análisis de gran calidad. Extraer datos para crear conjuntos de datos con la finalidad de desarrollar IA también genera conflictos de extracción y propiedad intelectual. Estos conjuntos, ya sean abiertos o cerrados, se suelen elaborar a partir de imágenes, texto y código de internet protegidos por copyright. Esto significa que, en la práctica, quienes utilizan estos conjuntos de datos para entrenar y evaluar modelos de IA a menudo se sirven de la propiedad intelectual y del trabajo ajenos para hacerlo, aduciendo un uso legítimo incluso cuando estos alegatos se basan en fundamentos jurídicos poco firmes. Sea legal o no, la práctica de arrasar indiscriminadamente con los datos de internet para crear sistemas que hoy por hoy logran debilitar el sustento de escritores, artistas y programadores —con cuyo esfuerzo se crearon esos datos de internet en primera instancia— ha generado alarma y enfado. Los sistemas de IA suelen ser contruidos por una minoría de empresas del mundo, con recursos de datos y de mano de obra obtenidos de la mayoría de la población mundial, a la que luego se imponen sin consulta previa.” (Whittaker, 2025, p.p-114-117).

Como bien explica Whittaker, *common crawl* es un “archivo masivo de datos de rastreo web recopilado por un pequeño grupo de personas que trabajan para una pequeña organización sin fines de lucro de California desde 2008” (Baack, 2004, p-2199). Estos archivos cuentan con alrededor de 9.5 *petabytes* de información; para tener un aproximado, un solo *petabyte* equivale a “un cuatrillón de *bytes*, o 1000 *terabytes*”(Lenovo, s. f.), esto se traduce en una cantidad masiva de datos disponibles, compuestos tanto por contenido de dominio público como por obras protegidas por *copyright*, todo ello accesible de manera prácticamente inmediata. A modo ilustrativo, la trilogía de El Señor de los Anillos en formato *epub* posee un peso aproximado de 4 MB, lo que permite dimensionar, aunque sea parcialmente, la magnitud de la información recopilada por plataformas como *common crawl*. Debido al enorme volumen de datos almacenados y procesados, este repositorio ha sido descrito en diversas ocasiones como una especie de copia de internet, aun cuando dicha expresión constituye claramente una hipérbole.

Es evidente que el acceso a esta información no es apta para todo el público, pues, un ciudadano común no posee la capacidad ni conocimiento para almacenar esta información, es por esto, que:

“los sistemas de IA no funcionan como el software tradicional; exigen procesos de desarrollo concretos y dependen de recursos especializados y onerosos que actualmente se acumulan en manos de unos pocos gigantes tecnológicos...Desarrollar poderosos sistemas de IA exige conjuntos de datos gigantescos que a su vez requieren una enorme capacidad computacional para poder ser procesados” (Whittaker, 2025, p.p -86-108).

Common Crawl es sumamente relevante y esencial para los LLMs, lo podemos ver reflejado en:

“GPT-3, el LLMs que impulsa la primera versión de Chat GPT publicada en 2022, no habría sido posible sin Common Crawl, ya que más del 80% de sus tokens¹ provenían de él. La mayoría de los LLMs publicados tanto antes como después por otros desarrolladores también dependen en gran medida de versiones filtradas de Common Crawl para su pre-entrenamiento.” (Baack, 2024, p- 2199).

Ahora bien, es importante resaltar lo siguiente expresado por Baack:

“A pesar de que la creación de estos archivos en 2007 no tenía fines de asistir al entrenamiento de sistemas de IA, hoy en día resulta un pilar fundamental para el entrenamiento de estos. Un breve inciso a este tema, es que las grandes empresas usan Common Crawl con cierta cautela, ya que, dentro de los archivos se encuentran materiales en cantidad significativa de contenido indeseable, incluyendo discursos de odio y contenido explícito” (Baack, 2024, p-2200)

Es decir, que la implementación de estos datos conlleva por parte de los programadores un previo análisis, conocido como Reinforcement Learning from Human Feedback (RLHF). En este proceso de RLHF Fenollosa explica su relevancia actual en las GenAI de la siguiente forma:

¹ “Los tokens de IA son pequeñas unidades de datos que surgen de descomponer grandes cantidades de información. Los modelos de IA procesan tokens para conocer las relaciones entre ellos y desbloquear capacidades que incluyen predicción, generación y razonamiento. Cuanto más rápido se puedan procesar los tokens, más rápido podrán aprender y responder los modelos” (Salvator, 2025). De forma resumida, toda la información recolectada es transformada en tokens para que los LLMs puedan digerir de forma más eficiente.

“Aprendizaje por refuerzo a partir de la supervisión humana. Actualmente, tutores están enseñando a Chat GPT a comportarse en sociedad usando interacciones previas de cómo comportarse, suministrando ejemplos e indicando lo que está bien y lo que está mal. No hay que confundir RLHF con pre programar las respuestas. El conocimiento siempre ha estado ahí, y el proceso de RLHF sencillamente desbloquea ciertas habilidades del modelo: sin RLHF, Chat GPT tendría muchas dificultades para hacer introspección sobre lo bueno y lo malo, lo conocido y lo desconocido, lo objetivo y la opinión, lo aceptable y lo inaceptable. La instrucción proporciona a los modelos dos características altamente deseables: la capacidad de resolver tareas nunca vistas, es decir, la generalización de problemas, y el alineamiento de las respuestas con las preguntas del usuario”(Fenollosa, 2024, p.p 80-81).

Es por esto, que las empresas como Open AI (Chat GPT) usan sus propios filtros a la hora de entrenar con datos, ya que, páginas como *Common Crawl* priorizan la cantidad más que la calidad.

Este fenómeno se desarrolla en un contexto caracterizado por un notable rezago normativo, donde extensos repositorios digitales que contienen material protegido por *copyright* -e incluso contenidos potencialmente sensibles- son utilizados de forma total para el entrenamiento de las GenAI. Mientras el usuario interactúa superficialmente con plataformas comerciales como ChatGPT para resolver dudas o ejecutar tareas mediante un *prompt*, permanece invisible una compleja infraestructura de recopilación, clasificación y procesamiento masivo de datos que opera en las etapas subyacentes del entrenamiento de los grandes LLMs.

Detrás de todo el proceso de alimentación de los LLMs para establecer una GenAI, existe un equipo de humanos que decide hacer caso omiso a legislaciones;

“dadas las limitaciones inherentes a la detección y eliminación automática de contenido problemático, es necesario establecer mejores estándares y prácticas en la industria para filtrar fuentes como Common Crawl que contienen dicho contenido. Si los creadores de LLMs insisten en utilizar dichas fuentes, es necesario debatir sobre la cantidad de contenido dañino tolerable, quién es el responsable y cómo abordar las consecuencias del entrenamiento de modelos con dicho contenido. Common Crawl hace que dichas directrices y mejores prácticas sean aún más urgentes porque su misión y, por consiguiente, sus datos no se alinean fácilmente con un desarrollo de LLMs justo y responsable”(Baack, 2024, p-2202).

En una entrevista al director de *Common Crawl* cuando se le preguntó sobre el material sensible en sus archivos, respondió lo siguiente:

“Desde el punto de vista de los objetivos, no creo que necesariamente queramos estar seleccionando cuidadosamente el conjunto de datos, ya que las páginas que eliminamos podrían ser valiosas para los usuarios posteriores. Podrías estar buscando la prevalencia del discurso de odio en un país determinado... Si eres el investigador que intenta medir la prevalencia, querrás que ese material esté ahí. Así que, en cierto modo, dijimos que la clasificación del contenido recae en el usuario final” (Baack, 2024, p- 2202).

Ahora bien, la protección de obras por *copyright* en el internet resulta complicado bajo el constante asecho de los *crawlers*. A pesar de que existen diferentes métodos que se usan para evitar que estos accedan a información que los titulares de los derechos no desean que sea usada para el entrenamiento de una IA, tales como: *Robots.txt.*, *NoAI meta tag* y *ai.txt*²; estos métodos resultan como ir a la guerra con un palo de madera; en el siguiente gráfico es visible la ausencia del respeto por parte de algunas empresas ante estos métodos.³

User Agent	Category	Company	Publish IP	Claim Respect	Respect in Practice
Amazonbot	AI Search	Amazon	Yes	Yes	Yes
AI2Bot	AI Data	Ai2	No	-	-
anthropic-ai	Undocumented AI	Anthropic	No	-	-
Applebot	AI Search	Apple	Yes	Yes	Yes
Applebot-Extended*	AI Data	Apple	-	Yes	-
Bytespider	AI Data	ByteDance	No	-	No
CCBot	AI Data	Common Crawl	Yes	Yes	Yes
ChatGPT-User	AI Assistant	OpenAI	Yes	Yes	Yes
Claude-Web	Undocumented AI	Anthropic	No	-	-
ClaudeBot	AI Data	Anthropic	No	Yes	Yes
cohere-ai	Undocumented AI	Cohere	No	-	-
Diffbot	AI Data	Diffbot	No	-	-
FacebookBot	AI Data	Meta	Yes	Yes	-
Google-Extended*	AI Data	Google	-	Yes	-
GPTBot	AI Data	OpenAI	Yes	Yes	Yes
Kangaroo Bot	AI Data	Kangaroo LLM	No	Yes	-
Meta-ExternalAgent	AI Data	Meta	Yes	-	Yes
Meta-ExternalFetcher	AI Assistant	Meta	Yes	No	-
OAI-SearchBot	AI Search	OpenAI	Yes	Yes	-
omgili	AI Data	Webz.io	No	Yes	-
PerplexityBot	AI Search	Perplexity	No	Yes	-
Timpibot	AI Data	Timpi	No	-	-
Webzio-Extended*	AI Data	Webz.io	-	Yes	-
YouBot	AI Search	You.com	No	-	-

(Enze, liu, 2025.p-81)

² Quien desee leer sobre cómo funciona cada una, recomiendo acceder al trabajo de Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. Enze Liu, Elisa Luo, Shawn Shan, Geoffrey M. Voelker, Ben Y. Zhao and Stefan Savage. Recuperado de: <https://dl.acm.org/doi/epdf/10.1145/3730567.3732913>

³ Somesite I Used To Crawl: Awareness, Agency and Efficacy in Protecting Content Creators From AI Crawlers. Enze Liu, Elisa Luo, Shawn Shan, Geoffrey M. Voelker, Ben Y. Zhao and Stefan Savage. Recuperado de: <https://dl.acm.org/doi/epdf/10.1145/3730567.3732913>. p 81)

El acceso a datos para el entrenamiento de las GenAI por parte de las empresas tecnológicas consisten -por lo menos aquellas en juicio en los Estados Unidos- en:

“el resultado de millones de artículos académicos, capítulos, libros y ponencias han sido agregados en colecciones ilícitas conocidas como “bibliotecas en la sombra” (shadow libraries), disponibles para su descarga gratuita por cualquier persona con una conexión a Internet. Las bibliotecas en la sombra más conocidas son Library Genesis (LibGen) y el sitio web asociado Sci-Hub, el cual utiliza credenciales universitarias compartidas para acceder a artículos e incorporarlos al repositorio de LibGen. AvaxHome y el desaparecido sitio web Library.nu son otros dos ejemplos, junto con numerosas colecciones más pequeñas dispersas por toda la red” (Gardner, 2017, p-1)

El entrenamiento de los sistemas desarrollados por empresas de GenAI implica, en numerosos casos, la utilización de grandes volúmenes de obras protegidas por *copyright*. Dentro de este contexto, las denominadas *shadow libraries* continúan siendo objeto de un debate jurídico en lo relativo a su impacto sobre el mercado y el desplazamiento de las obras protegidas. La ausencia de consenso respecto a los efectos económicos derivados de estos repositorios digitales puede observarse claramente en los litigios promovidos por los titulares de derechos contra empresas tecnológicas.

En este orden de ideas, y cuando las primeras empresas de GenAI comenzaron a recolectar datos de forma masivos para entrenar sus modelos, no inventaron una nueva forma de obtener obras protegidas: heredaron y aprovecharon la infraestructura tecnológica de las *shadow libraries* que la academia y la industria editorial conocían desde hacía años. *LibGen* y sus repositorios asociados ofrecían precisamente lo que un desarrollador de LLMs necesitaba: decenas de millones de obras digitalizadas, descargables masivamente, sin barreras técnicas ni costo aparente. Lo que la litigación reciente ha hecho visible es que esta infraestructura, antes utilizada principalmente por investigadores individuales en busca de artículos específicos, fue empleada en escala industrial por empresas comerciales con valoraciones multimillonarias para construir productos comerciales también multimillonarios. En este punto se concentra el núcleo problemático que da origen al presente TEG.

La diferencia fundamental entre los *crawlers* y las *shadow libraries* -siendo estas últimas el eje central de los litigios actuales en EE.UU- radica en su alcance y propósito. Mientras que los primeros recopilan de manera automatizada todo tipo de información disponible en la web, indexando tanto contenidos de dominio público como privado, las *shadow libraries* se dedican exclusivamente a la compilación y distribución masiva de obras protegidas. Este fenómeno emula, en el ámbito estrictamente literario y académico, el modelo de plataformas de intercambio no autorizado como The Pirate Bay, resultando en un insumo altamente provechoso para el entrenamiento de los sistemas de GenAI.

Copyright, machine learning y minería de texto y datos.

El ML es aquella capacidad de los sistemas de IA para resolver problemas reales tratando de emular el cerebro humano mediante el análisis de datos; una definición suministrada por Izquierdo:

“ML es el nombre de una familia integral de técnicas utilizadas para nuevas formas de análisis de datos de IA que se convierten en herramientas esenciales para la predicción y la toma de decisiones. Existe una mayor capacidad para entrenar algoritmos existentes con grandes cantidades de muestras de datos.”(Izquierdo, 2022, p-505).

Es por esto que los datos y el acceso a esta información por parte de las empresas tecnológicas de sistemas de IA, resultan de suma importancia, llegando a la célebre frase del matemático británico Clive Humby “los datos son el nuevo petróleo” (University of Sheffield,s/f), debido a su importancia en el mundo globalizado por la tecnología. Para tratar de aproximarnos a esta relación entre datos e IA, Rouhiainen establece:

“Desde el punto de vista del desarrollo, una razón obvia del crecimiento de la IA es que el poder de procesamiento de los ordenadores ha aumentado exponencialmente, lo que les permite procesar más algoritmos complejos. Éstos son los mismos tipos de algoritmos avanzados que mueven la IA. Por otra parte, los datos son otro elemento importante para el desarrollo de la IA: sin ellos, sería casi imposible crear productos y aplicaciones con esta tecnología” (Rouhiainen, 2018, p-25)

Todo esto se desenlaza en que:

“La IA tiene un enorme apetito por los datos. Entre más datos, mejor. Sus cálculos se vuelven cada vez más precisos. Por lo cual, las redes neuronales necesitan un acceso cada vez más grande a volúmenes de información (datos), incluyendo una cantidad enorme de trabajos protegidos por derechos de propiedad intelectual.” (Anguiano,2023, p-63)

Incluso, Rouhiainen considera esta materia de la siguiente forma:

“Sin embargo, personalmente, yo diría que los datos son incluso mejores que el petróleo. En los años en los que éste era uno de los productos básicos más valiosos del mundo, sólo pocas empresas podían disfrutar de sus beneficios. No obstante, hoy día, debido a que casi cualquier persona puede aprender los principios básicos de la IA y el aprendizaje automático, usar estas habilidades para crear herramientas valiosas y tener acceso a fuentes de datos gratuitos online con facilidad supone un nuevo escenario en el que todos podrán beneficiarse de ello.(L. Rouhiainen, 2018, p-25)⁴

Es por esto que surge la necesidad del TDM, la cual se define como:

⁴ Aquí difiero un poco con el autor, ya que, como mencione anteriormente, albergar un sistema de GenAI amerita de una estructura inmensa y costosa.

“Los algoritmos de TDM están diseñados para analizar grandes cantidades de datos con el fin de identificar patrones y correlaciones que puedan, ya sea directa o indirectamente, ayudar a explicar las relaciones causales asociadas con los fenómenos naturales bajo investigación... La promesa de estas tecnologías es que pueden ser capaces de identificar patrones que, de otro modo, sólo surgirían tras años de experimentación de ensayo y error, o que podrían no llegar a reconocerse nunca. En su forma más ambiciosa, una tecnología de TDM utilizaría el procesamiento semántico para ser capaz de analizar estos patrones y emitir juicios o conclusiones tan buenos o mejores que los de un investigador experto sobre el significado y la importancia de estos hallazgos. A más corto plazo, estas tecnologías pueden proporcionar aportes fundamentales sobre la manera en que los investigadores formulan sus preguntas de investigación. Por ejemplo, si un investigador busca comprender el papel de un gen específico o un conjunto de genes en relación con la trayectoria de una enfermedad, podría ejecutar un análisis de todos los artículos que informan resultados de experimentos de cualquier tipo relacionados con dicho gen o genes. Los resultados de este análisis podrían identificar una correlación inesperada entre el gen (o los genes) y alguna otra parte del cuerpo que, de otro modo, parecería no estar relacionada con la enfermedad bajo estudio. Este hallazgo probablemente llevaría al investigador a comprobar si existe alguna relación. En el mejor de los casos, los resultados demostrarían la existencia de dicha relación y cómo funciona, lo que desbloquearía una solución que había estado frenando a los investigadores en el desarrollo de un compuesto o terapia génica relevante” (Caroll, 2019, p.p - 901- 902).

Aunque estas ventajas son evidentes en la sociedad actual, y el TDM ofrece un soporte elemental para el avance científico, debemos irnos de lo macro a lo micro y direccionarnos a las GenAI, en donde dicha función no resulta tan filantrópica, pues, ya no hablamos de científicos para determinar o encontrar alguna cura para enfermedades, nos referimos a aquellas prácticas que solo se han convertido mayormente en mero espectáculo y entretenimiento como lo son Chat GPT, Gemini o Claude; en donde el TDM y el RLHF se retroalimentan con el fin de que el *output* de estas Gen AI sean lo más similares a un humano posible, además, y como se mencionó anteriormente, sirven para filtrar aquellos datos/tokens que sean potencialmente dañinos.

Un ejemplo ilustrativo para entender todo este proceso y las diferencias entre *crawling*, TDM y RLHF, sería el siguiente: Los *crawlers* son aquellos que van a buscar la materia prima; buscan la piedra en bruto. El TDM es el proceso de refinación para obtener los minerales deseados con los que poder trabajar, sin embargo, se somete a un doble chequeo por parte de un supervisor con el fin de asegurar que sea el mineral deseado comerciable.

Ahora bien, resulta indispensable analizar este fenómeno desde una perspectiva multidimensional que permita identificar con precisión los beneficios y los perjuicios de estas tecnologías. La demanda de insumos para el adiestramiento de los sistemas de IA evidencia que los datos de obras en dominio público resultan insuficientes para las necesidades de optimización de los modelos actuales de GenAI. Como consecuencia de este requerimiento expansivo de información, las empresas desarrolladoras incorporan de forma sistemática obras protegidas por el copyright. Si bien el ordenamiento de EE.UU prevé la doctrina del *fair use* para amparar ciertos usos, resulta imperativo examinar sus posibles desnaturalizaciones y determinar el núcleo de la controversia.

El presente TEG no consiste en el misoneísmo, no busca rechazar los avances tecnológicos y los beneficios de estos sistemas de IA generales; se entiende los grandes aportes de estas, tales como: “El TDM se puede utilizar en lingüística para analizar grandes cuerpos de texto y extraer patrones sintácticos o gramaticales; en medicina, puede encontrar asociaciones entre un gen y una enfermedad, o entre un fármaco y un evento adverso; En general, el TDM se está convirtiendo en la práctica estándar en el sector farmacéutico” (Izquierdo, pp-498).

Como se esbozó con anterioridad, el objeto de análisis se centra en corporaciones como OpenAI y Anthropic; cuya alimentación y perfeccionamiento derivan de una cadena de procesos donde la supervisión humana desempeña un papel fundamental..

Bases legales

Copyright Act

El *Copyright Act* de 1976 con sus correspondientes enmiendas a lo largo de los años, es la actual legislación que se aplica en EE.UU. En su sección §102 establece lo siguiente:

“(a) La protección de copyright subsiste, de conformidad con este título, en obras originales de autoría fijadas en cualquier medio tangible de expresión, ya sea conocido ahora o desarrollado posteriormente, desde el cual puedan ser percibidas, reproducidas o comunicadas de otro modo, ya sea directamente o con la ayuda de una máquina o dispositivo. Las obras de autoría comprenden las siguientes categorías: obras literarias; obras musicales, incluyendo cualquier letra que las acompañe; obras dramáticas, incluyendo cualquier música que las acompañe; pantomimas y obras coreográficas; obras pictóricas, gráficas y escultóricas; películas cinematográficas y otras obras audiovisuales; grabaciones sonoras; y obras arquitectónicas. (b) En ningún caso la protección de copyright para una obra original de autoría se extiende a cualquier idea, procedimiento, proceso, sistema, método de operación, concepto, principio o descubrimiento, independientemente de la forma en que sea descrito, explicado, ilustrado o incorporado en dicha obra” (U.S. Copyright Office, 2021, § 102).

Andrea Bartz, Charles Graeber, and Kirk Wallace Johnson v. Anthropic pbc,

Uno de los precedentes fundamentales para ilustrar la dimensión fáctica de los conceptos y procesos analizados a lo largo de este TEG, es el litigio Bartz et al. v. Anthropic PBC. De manera específica, resulta de gran relevancia el sumario judicial emitido por el Tribunal de Distrito de EE.UU para el Distrito Norte de California, con fecha del 23 de junio de 2025. La introducción de dicho documento establece lo siguiente:

“Una empresa de IA descargó de forma gratuita millones de libros protegidos por derechos de autor en formato digital desde sitios piratas en internet. La empresa también compró libros protegidos (algunos de los cuales coincidían con los adquiridos en los sitios piratas), les quitó la encuadernación, escaneó cada página y los almacenó en archivos digitalizados en los que se pueden realizar búsquedas. Todo lo anterior se realizó para acumular una biblioteca central de 'todos los libros del mundo' para conservarlos 'por siempre'. A partir de esta biblioteca central, la empresa de IA seleccionó diversos conjuntos y subconjuntos de libros digitalizados para entrenar varios modelos de lenguaje de gran escala en desarrollo para potenciar sus servicios de IA. Algunos de estos libros fueron escritos por los autores demandantes, quienes ahora demandan por infracción de copyright. En el juicio sumario, la cuestión es hasta qué punto cualquiera de los usos de las obras en cuestión califica como 'uso legítimo' (fair use) bajo la Sección 107 de la Copyright Act "(Bartz v. Anthropic, 2025).

Ahora bien, la empresa demandada es Anthropic, la cual se define en su página web como “una compañía de investigación y seguridad en IA. Desarrollamos sistemas fiables, interpretables y controlables” (Anthropic, s.f. párr. 5). Se trata de una empresa dedicada al desarrollo de GenAI y creadora de Claude, su propio LLM. De manera similar a como OpenAI desarrolla ChatGPT o Google desarrolla Gemini, Anthropic ha diseñado Claude como uno de sus principales sistemas de GenAI.

Claude fue lanzado al mercado en marzo de 2023 por Anthropic, empresa fundada por antiguos empleados de OpenAI. Al igual que otros sistemas de GenAI, Claude dispone de una versión gratuita; no obstante, también ofrece planes de suscripción pagos. La versión *Pro* tiene un costo aproximado de diecisiete dólares estadounidenses (US\$17) mensuales, mientras que la versión *Max* alcanza los cien dólares estadounidenses (US\$100) por mes.

Anthropic para el entrenamiento de su LLM usó diferentes obras de los demandantes sin su consentimiento, en específico, copió las siguientes obras de los autores que mencionan la apropiación sin su consentimiento de: *The Lost Night: A Novel*, *The Herd*, *We Were Never Here* y *The Spare Room*, de Andrea Bartz; *The Good Nurse: A True Story of Medicine, Madness, and Murder*, and *The Breakthrough: Immunotherapy and the Race to Cure Cancer*, de Charles Graeber; y, *To Be A Friend Is Fatal: The Fight to Save the Iraqis America Left Behind*, *The Feather Thief: Beauty, Obsession, and the Natural History Heist of the Century*, and *The Fishermen and the Dragon: Fear, Greed, and a Fight for Justice on the Gulf Coast*, de Kirk Wallace Johnson (Bartz v. Anthropic, 2025). Estos libro, entre novelas y obras de no ficción oscilan sus precios de un dólar americano (\$1) hasta veintinueve dólares con sesenta y tres centavos (\$29, 63); todo depende del formato en el que se desee el libro y son fácilmente accesibles en Amazon y *Kindle Unlimited*.

Ahora bien, antes de analizar el caso a mayor profundidad, es preciso determinar el criterio adoptado por William Alsup, juez a cargo de la causa. El juzgado realiza una nítida distinción entre el material adquirido por Anthropic de forma lícita y aquel proveniente de fuentes no autorizadas o ilícitas. Con respecto a la primera, establece: “esta orden concede el juicio sumario a favor de Anthropic, dictaminando que el uso para entrenamiento constituyó un *fair use*” (Bartz v. Anthropic, 2025). Es importante destacar que todo el caso se basa en la figura del *fair use* con respecto al entrenamiento del LLM de dicha empresa con el fin de alimentar a su GenAI llamada Claude. Por otro lado, declara:

“Se llevará a cabo un juicio sobre las copias piratas utilizadas para crear la biblioteca central de Anthropic y los daños resultantes, ya sean reales o establecidos por la ley (incluyendo aquellos por intencionalidad). El hecho de que Anthropic comprara posteriormente una copia de un libro que previamente robó de internet no la eximirá de responsabilidad por la copia ilícita, aunque podría afectar la cuantía de los daños legales. No se excluye ninguna acción respecto a otras copias derivadas de la biblioteca para usos distintos al entrenamiento de LLMs (Bartz v. Anthropic, 2025).

Expuesto lo anterior, corresponde entonces determinar cuál es el problema jurídico central planteado en el presente caso: “En agosto de 2024, los tres autores mencionados interpusieron esta demanda colectiva presuntiva, alegando que Anthropic había infringido sus derechos de autor federales al piratear copias para su biblioteca y al reproducirlas para entrenar su LLM” (Bartz v. Anthropic, 2025). Como ya se ha mencionado en el marco teórico, las empresa como Anthropic necesitan alimentar a sus GenAI con una cantidad abrumadora de datos, y como se establece en el caso, la empresa ha creado bibliotecas personales con una cantidad considerable de datos, donde Anthropic declaró en un inicio:

“La piratería de las copias iniciales de los libros de los autores y de millones de otros libros estaba justificada porque todas esas copias eran, al menos, razonablemente necesarias para el entrenamiento de LLMs y, sin embargo, Anthropic se ha resistido a incluir en el expediente qué copias o incluso qué conjuntos de copias se utilizaron de hecho para dicho entrenamiento” (Bartz v. Anthropic, 2025).

Dicho esto, Anthropic creó una biblioteca de investigación general con la finalidad de tener acceso a una gran cantidad de datos que ayudarían a “alimentar y mejorar a su LLM para ofrecer un mejor servicio pago para los usuarios de Claude” (Bartz v. Anthropic, 2025). Ahora bien, ¿cuál era el propósito de alimentar a esta GenIA con obras de ficción? Pues, la respuesta es simple: “Anthropic llegó a valorar por encima de todo, para sus mezclas de datos, libros como los que los Autores habían escrito, y los valoraba precisamente por las expresiones creativas que contenían” (Bartz v. Anthropic, 2025).

En relación con lo expuesto a lo largo del presente TEG, Anthropic sostiene que, mediante sus procesos de TDM y las posteriores técnicas de RLHF, los resultados generados por Claude se encuentran programados para evitar la reproducción textual de las obras utilizadas durante el entrenamiento del modelo. En este sentido, los demandantes no alegan ni han logrado comprobar que el sistema proporcione copias exactas de los libros presuntamente utilizados para entrenar el LLM de Anthropic, razón por la cual dicha cuestión queda excluida del objeto principal de análisis del presente caso.

No obstante, el hecho de que no se haya demostrado por los momentos que Claude no reproduce copias exactas en sus *outputs*, no elimina el problema relativo a la conservación interna de las obras utilizadas por Anthropic. Según consta en el expediente, “las copias se mantenían en la versión original subyacente de los archivos de libros que la empresa había obtenido o creado, ya fuera mediante descarga de fuentes ilícitas o mediante escaneo de ejemplares físicos”(Bartz v. Anthropic, 2025). Asimismo, Anthropic habría planeado conservar dicha biblioteca de forma indefinida, bajo el argumento de que no existía una razón suficiente para eliminar un libro, incluso si este no era utilizado directamente para entrenar a su LLM. Con el tiempo, la empresa también invirtió en el desarrollo de herramientas destinadas a facilitar búsquedas dentro de esa biblioteca de propósito general y a permitir el acceso a libros individuales o conjuntos de libros para usos adicionales.

Dicho esto, corresponde observar el análisis del juez, empezando por la Sección 107 del Copyright Act el cual establece los numerales del *fair use*:

“(1) El propósito y el carácter del uso, incluyendo si dicho uso es de naturaleza comercial o tiene fines educativos sin ánimo de lucro; (2) la naturaleza de la obra protegida por copyright; (3) la cantidad y la importancia de la parte utilizada en relación con la totalidad de la obra protegida; y (4) el efecto del uso sobre el mercado potencial o el valor de la obra protegida” (U.S. Copyright Office, 2025).

Con respecto al primer punto del *fair use*, el juez establece que el uso de las copias usadas por parte de Anthropic para entrenar a sus LLMs resulta increíblemente transformador, además, establece la siguiente comparación:

“Los Autores no pueden excluir legítimamente a nadie del uso de sus obras para formación o aprendizaje como tal. Todo el mundo también lee textos y luego escribe textos nuevos. Puede que necesiten pagar para tener en sus manos el texto. Pero hacer pagar a cualquiera específicamente para el uso de un libro cada vez que lo leen, cada vez que lo recuerdan de memoria, cada vez más tarde recurren a él al escribir cosas nuevas sería impensable. Durante siglos hemos leído y releído libros. Hemos admirado, memorizado e internalizado sus amplios temas, sus puntos sustantivos y sus soluciones estilísticas a la escritura recurrente problemas” (Bartz v. Anthropic, 2025).

El razonamiento expuesto por el juez resulta acertado hasta cierto punto, especialmente si se considera que el aprendizaje humano históricamente se ha desarrollado mediante la lectura, interpretación y análisis de textos. En este sentido, uno de los propósitos fundamentales del sistema de *copyright* no consiste en obstaculizar la educación, el intercambio de ideas o el progreso intelectual. No obstante, trasladar automáticamente dicha lógica al funcionamiento de sistemas de GenAI plantea importantes dificultades jurídicas y conceptuales, particularmente porque modelos como Claude no constituyen seres humanos, sino herramientas tecnológicas desarrolladas con fines comerciales. En consecuencia, la utilización de analogías entre el aprendizaje humano y el entrenamiento de grandes LLMs evidencia las limitaciones actuales del marco normativo frente a tecnologías emergentes y el uso masivo de obras protegidas durante el entrenamiento de sistemas de IA. Adicionalmente el juez menciona :

“En resumen, el propósito y el carácter de utilizar obras protegidas por copyright para entrenar LLMs con el fin de generar texto nuevo fue quintaesencialmente transformador. Al igual que cualquier lector que aspira a ser escritor, los LLMs de Anthropic se entrenaron con estas obras no para tomar ventaja y replicarlas o suplantarlas, sino para dar un giro radical y crear algo diferente. Si este proceso de entrenamiento requirió razonablemente realizar copias dentro del LLMs o de otra manera, dichas copias formaron parte de un uso transformador.”(Bartz v. Anthropic, 2025)

No obstante, uno de los aspectos más relevantes de la decisión analizada radica en la utilización del concepto de “quintaesencialmente transformador” para describir el entrenamiento de los LLMs. En términos técnicos, los LLMs funcionan mediante procesos de tokenización, a través de los cuales los datos utilizados durante el entrenamiento son convertidos en unidades matemáticas denominadas tokens. Tal como se expone en el sumario analizado, “algunas palabras fueron comprimidas en formatos más reducidos, por ejemplo: studying a study. Y todos los caracteres se agrupan en secuencias cortas y se tradujeron al número correspondiente de secuencias o tokens según un diccionario elaborado por Anthropic” (Bartz v. Anthropic, 2025).

Estos tokens no son más que la forma matemática para poder enseñar a los LLMs, con el fin de no crear algo nuevo, sino predecir. Estos sistemas se basan en estadísticas matemáticas con respecto a los tokens, “el aprendizaje profundo no es razonamiento o inteligencia sino estadística” (Fenollosa, 2024, p-99).

Para este punto es fundamental mencionar a Emily Bender y su trabajo titulado Loro Estocástico, el cual establece:

"Al contrario de lo que pueda parecer cuando observamos su resultado (output), un modelo de lenguaje es un sistema para coser al azar secuencias de formas lingüísticas que ha observado en sus vastos datos de entrenamiento, de acuerdo con información probabilística sobre cómo se combinan, pero sin ninguna referencia al significado: un loro estocástico" (Bender, 2021, pp 617).

A pesar de que este argumento de Bender suscita objeciones válidas, resulta indudable que plantea la necesaria distinción entre el ser humano y la IA. En la actualidad, esta tecnología no replica la conciencia ni el razonamiento humano; por ende, ante la falta de una definición conceptual unívoca de la IA y la persistente ausencia de regulaciones legales, resulta inviable aplicar dicha analogía para otorgarle un trato equivalente al de una persona; La Fontaine comenta al respecto:

"Estas tecnologías aún enfrentan desafíos significativos, como la incapacidad para realizar razonamientos complejos a nivel humano y la comprensión de contextos más profundos. La capacidad de los modelos de lenguaje para simular conversaciones y generar respuestas coherentes no debe confundirse con una auténtica inteligencia o consciencia" (La Fontaine, 2024, p-85).

Con respecto al almacenamiento de los datos en librerías privadas, el juez establece que para aquellas obtenidas de forma legal no ocasionan problemas y se consideran dentro del *fair use* al no tener pruebas de que Anthropic haya compartido o vendido estas con otras personas o empresas, sin embargo, para las copias piratas se menciona:

"Las bibliotecas universitarias y Google hicieron esfuerzos extraordinariamente grandes para garantizar que todas las copias estuvieran protegidas contra usos no autorizados, tanto mediante medidas técnicas como a través de acuerdos legales entre todos los participantes. No ocurre lo mismo aquí. Las copias de las bibliotecas carecían de controles internos que limitaran el acceso y el uso" (Bartz v. Anthropic, 2025).

Lo anterior permite inferir que los repositorios creados por Anthropic carecen de mecanismos adecuados de control y protección sobre el material almacenado. En consecuencia, existe la posibilidad de que terceros, mediante técnicas de *crawling* o procesos de TDM, accedan a obras obtenidas de manera ilícita o almacenadas sin autorización, circunstancia que fortalece los argumentos relativos a la posible responsabilidad de Anthropic respecto al manejo y utilización de dicho contenido protegido.

Respecto al segundo factor del *fair use* relativo a la naturaleza de la obra protegida, el Tribunal resolvió en contra de la parte demandada. Ello se debe a que el material utilizado durante el entrenamiento no constituía un conjunto de datos aleatorio, sino que respondía a una intención deliberada por parte de Anthropic. En efecto, se determinó que "las obras fueron elegidas por sus cualidades expresivas para la construcción de una biblioteca central y, posteriormente, para el entrenamiento de LLMs específicos" (Bartz v. Anthropic, 2025). En consecuencia, al comprobarse que el valor creativo y expresivo de los textos constituía precisamente el interés para el entrenamiento del modelo, este factor no pudo ser considerado como un uso inocuo ni plenamente justificable bajo la doctrina del *fair use*.

El tercer factor con respecto a la cantidad y la importancia de la parte utilizada en relación con la totalidad de la obra protegida, “el punto clave de este factor es si la cantidad fue razonable en relación con el propósito de la copia”(Bartz v. Anthropic, 2025); por lo tanto a las copias obtenidas de forma legal, el juez establece que “no existe ninguna alegación de una conexión rastreable entre las respuestas (outputs) del servicio Claude y las obras de los autores. Por lo tanto, el copiado utilizado para entrenar los LLMs que sustentan a Claude fue especialmente razonable" (Bartz v. Anthropic, 2025)". No obstante, al realizar una prueba exploratoria mediante un *prompt* a Claude, el sistema es capaz de generar un análisis detallado de la obra, incluyendo elementos como la trama principal y determinados giros argumentales. Si bien este ejercicio no constituye por sí mismo una prueba concluyente dentro del litigio, sí permite ilustrar las capacidades actuales de estos modelos y evidencia la necesidad de continuar examinando con mayor profundidad el alcance de este factor dentro del análisis del *fair use*.

Por último, corresponde analizar el cuarto factor del *fair use*: el efecto del uso sobre el mercado potencial de la obra protegida o, en otros términos, la posibilidad de desplazamiento de compradores potenciales mediante el uso del modelo Claude. Este análisis fue abordado por el Tribunal distinguiendo entre las copias obtenidas de forma lícita y aquellas provenientes de fuentes ilícitas.

Respecto de las obras adquiridas legalmente, ante el argumento de los demandantes según el cual Claude podría convertirse en un competidor dentro del mercado creativo y editorial, el juez sostuvo que:

“La demanda de los autores no es diferente de lo que sería si se quejan de que enseñar a los niños a escribir bien resultaría en una explosión de obras competidoras. Este no es el tipo de desplazamiento competitivo o creativo que preocupa al Copyright Act. La Ley busca promover las obras originales, no proteger a los autores contra la competencia” (Bartz v. Anthropic, 2025).

A pesar de la fuerza argumentativa de dicha afirmación, el razonamiento del Tribunal puede considerarse problemático desde una perspectiva conceptual. Equiparar el entrenamiento de un sistema de GenAI con el proceso cognitivo y formativo de un niño implica trasladar categorías humanas a estructuras tecnológicas que operan bajo principios radicalmente distintos. La creatividad constituye una facultad inherente a la experiencia humana, mientras que modelos como Claude funcionan mediante procesos estadísticos y computacionales sustentados en redes neuronales artificiales y grandes volúmenes de datos. En consecuencia, la ausencia de regulación específica aplicable a estas tecnologías no debería conducir automáticamente a su equiparación analógica con el aprendizaje humano, especialmente cuando detrás de estos sistemas existen corporaciones con claros intereses económicos y competitivos.

Por otra parte, en relación con las copias obtenidas mediante fuentes ilícitas, el Tribunal resolvió en contra del *fair use*, al considerar que Anthropic incorporó obras provenientes de plataformas de piratería masiva que razonablemente pudieron haber sido adquiridas a través de canales comerciales legítimos.

Dicho todo esto, el Tribunal llega a la conclusión de un uso transformativo y del *fair use* sobre las obras adquiridas de forma legal, mientras que las obtenidas de forma ilícita serán llevadas a un juicio más adelante; quedará ver cómo los tribunales Estadounidenses deciden resolver dicha disputa, ya que, como declara la empresa Anthropic: "las copias de las obras (incluidas las piratas) se conservarán para siempre para fines generales, incluso después de que Anthropic determinara que nunca se utilizarían para el entrenamiento de los LLMs" (Bartz v. Anthropic, 2025). Esto último representa un riesgo enorme debido a la posible falta de seguridad y facilidad para acceder a estas bibliotecas digitales debido a los *crawlers* y TDM.

New York Times: noticias en el alimento de inteligencia artificial generativa

El 27 de diciembre de 2023, The New York Times Company (NYT) interpuso ante el Southern District of New York una demanda contra Microsoft Corporation y OpenAI que cambió, de manera definitiva, la escala del conflicto. Hasta ese momento, los demandantes eran autores individuales o grupos pequeños representados por despachos boutique. Con la entrada del NYT, el litigio dejó de ser una disputa entre autores aislados y empresas tecnológicas para convertirse en un enfrentamiento entre una de las instituciones periodísticas más grandes de Estados Unidos y dos de las corporaciones más valiosas del mundo. La demanda del NYT, además, introducía una novedad analítica: no se limitaba a sostener que los artículos habían sido utilizados para el entrenamiento; sostenía, con evidencia técnica obtenida antes de la presentación, que los outputs de ChatGPT funcionaban como sustitutos económicos directos del periodismo original (Columbia Undergraduate Law Review, 2025; NPR, 2025. párr. 12).

“Los modelos más recientes de GPT están entrenados con billones de palabras; un conjunto de datos tan grande que sería el equivalente a un documento de Microsoft Word de más de 3,700 millones de páginas. La raíz de la demanda del Times es que este conjunto de datos contiene una masa de contenido del Times protegido por copyright... La alegación central del Times es que OpenAI está infringiendo los derechos de autor mediante el uso y la reproducción no autorizados y sin licencia de las obras del periódico durante el entrenamiento de sus modelos. Pero el problema se amplifica de dos maneras. Primero, los LLMs a veces "memorizan" partes de las obras incluidas en los datos de entrenamiento. Cuando esto sucede, los modelos pueden generar ocasionalmente reproducciones casi literales de las obras. Segundo, y relacionado con lo anterior, los LLMs producen resultados de búsqueda "sintéticos" que, bajo petición, pueden reproducir "contenido significativamente más expresivo de [un] artículo original de lo que tradicionalmente mostraría" una búsqueda en línea, permitiendo efectivamente que los lectores evadan el muro de pago (paywall) del Times” (Pope, 2024.párr.6).

El núcleo de esta demanda se ha explicado a lo largo del presente TEG, proceso en el que vemos desde los *crawlers*, TDM, RLHF y finalizando con el entendimiento de que un sistema de GenAI se basa; en palabras de J, Astudillo:

“La profundidad que obtengo es, en buena medida, una función de la capacidad estadística de la máquina para imitar textos profundos: hay citas, matices, un registro que suena a ensayo serio. La pregunta es si, en esas dos horas, mi propio pensamiento se ha vuelto más profundo o si, simplemente, me he apoyado en un simulacro de profundidad producido a gran velocidad.”(Astudillo, 2026, p- 6)

Ahora bien, la presentación del NYT desencadenó, a lo largo de 2024, una cascada de demandas paralelas por parte de otros medios: Daily News, Center for Investigative Reporting, Raw Story Media, AlterNet, The Intercept. La mayoría de estas causas terminaron consolidadas, en septiembre de 2024, en lo que se conoce hoy como *In re OpenAI, Inc. Copyright Infringement Litigation*, un multidistrito (MDL) que aglutina los principales pleitos periodísticos contra OpenAI ante la jueza Sidney Stein (S,Rao;I,Estoesta, 2026. párr. 21)

Estos casos siguen en pie y todavía hasta la presente fecha de este TEG no se ha llegado a una conclusión, por lo cual, tendremos que estar pendiente de los desenlaces por venir en los siguientes meses.

Kadrey et al. v. Meta Platforms

Kadrey et al. v. Meta Platforms, Inc es de los casos de más larga data en los tribunales estadounidenses. Presentado en julio de 2023 por trece autores —incluyendo Richard Kadrey, Sarah Silverman y Ta-Nehisi Coates— el caso alega que Meta utilizó sus obras para entrenar los modelos Llama mediante entrenamiento directo desde librerías piratas digitales con conocimiento y autorización explícita de la dirección de la empresa, incluyendo la del propio Mark Zuckerberg (TechCrunch, 2025. párr. 8).

El Juez Vince Chhabria dictó el 25 de junio de 2025 juicio sumario a favor de Meta respecto a las reclamaciones principales, con base en insuficiencia probatoria: “los demandantes no demostraron que el modelo Llama reprodujeran sus obras ni que los *outputs* del modelo causan daño económico concreto al mercado de sus publicaciones” (Kadrey v. Meta Platforms, Inc., 2025).

Sin embargo, el Tribunal rechazó la *Shadow Library Strategy*⁵ en Kadrey con un razonamiento opuesto al del Juez Alsup en Bartz: el Juez Chhabria consideró que la adquisición de copias desde *shadow libraries* era parte integral del propósito transformativo de entrenamiento, no una conducta separable (Lee, 2025.párr. 4).

Esta divergencia entre Alsup y Chhabria sobre las *Shadow Library Strategy* es la tensión doctrinal más relevante del litigio literario en EE.UU a la fecha del presente TEG. Dos jueces del mismo Distrito, resolviendo hechos estructuralmente similares, llegaron a conclusiones opuestas sobre la separabilidad jurídica de la adquisición y el entrenamiento. La resolución de estas divergencias todavía está por decidirse.

El caso Kadrey v. Meta Platforms tampoco ha concluido. Los demandantes continuarán con la demanda por la piratería de sus obras en el entrenamiento de las GenAI y solicitarán nueva determinación de juicio.

CAPÍTULO III: MARCO METODOLÓGICO

Línea de investigación

Copyright en el sistema de los EE.UU: El objeto del presente TEG consiste en enfocar el impacto y avance de las nuevas tecnologías como lo son las GenAI y tratar de desglosar sus funcionamientos, métodos de alimentación y aprendizaje, con el fin de identificar aquellas zonas donde existen fricciones legales con respecto a los titulares de los derechos.

⁵ La *Shadow Libraries Strategy* es un término acuñado por Edward Lee para nombrar una línea de defensa en los juicios de GenAI y *copyright*; en el cual los demandantes afinan sus defensas en la alimentación mediante páginas web con repositorio de obras obtenidas de manera ilícitas.

Tipo de investigación

El presente TEG se basa en investigación documental, la cual consiste en “un procedimiento científico, un proceso sistemático de indagación, recolección, organización, análisis e interpretación de información o datos en torno a un determinado tema. Al igual que otros tipos de investigación, éste es conducente a la construcción de conocimientos” (Alfonso, 1995, p-2).

Adicionalmente, “la investigación documental tiene la particularidad de utilizar como una fuente primaria de insumos, más no la única y exclusiva, el documento escrito en sus diferentes formas: documentos impresos, electrónicos y audiovisuales” (Morales, 2003, p-2).

Dicho esto, el presente TEG posee un carácter eminentemente documental, en tanto se fundamenta en la recopilación y análisis de libros, artículos académicos, *papers* y decisiones judiciales en EE.UU orientadas a comprender la situación actual relacionada con el entrenamiento de los sistemas GenAI. No obstante, la investigación no pretende ofrecer una solución definitiva al problema planteado, debido a que gran parte de los litigios y debates jurídicos sobre esta materia aún se encuentran en pleno desarrollo.

Diseño de la investigación

En esta línea de trabajo se ha usado la investigación cualitativa, la cual se entiende de la siguiente forma:

“La Investigación Cualitativa es en esencia un multimétodo, involucra un enfoque naturalista, interpretativo para sus objetos de estudio. Esto significa que los investigadores cualitativos estudian cosas en sus escenarios naturales, intentan obtener el sentido de, o interpretar, los fenómenos en términos de los significados que las personas les asignan. La investigación cualitativa involucra el uso intencional y la obtención de una variedad de materiales empíricos –estudio de caso, experiencia personal, introspección, historia de vida, entrevista, observación, lo histórico, interacciones y textos visuales– descritos como momentos rutinarios y problemáticos, y significativos en las vidas de los individuos” (Macía, 2023, p-11).

Técnicas y herramientas de recolección y procesamiento de datos

Para el desarrollo del presente TEG, se implementó la recolección y procesamiento de la información mediante la técnica de revisión documental. “Esta técnica consiste en detectar, obtener y consultar bibliografía y otros materiales útiles para los propósitos del estudio, así como en extraer y recopilar la información relevante que atañe al problema de investigación” (Hernández Sampieri et al., 1991, p-26). En ese sentido, se realizó una compilación y crítica de fuentes normativas, doctrinales y jurisprudenciales vinculadas al objeto de estudio, confrontando el contenido de cada documento a fin de garantizar la fiabilidad y precisión de la información analizada.

Presentación y análisis de resultados

Tras la investigación documental de los diferentes casos todavía en juicio como de los diferentes autores que tratan de aportar su perspectiva en este nuevo tema en auge, se llega a la conclusión de que actualmente no existe un consenso y se da a entrever las divergencias interpretativas existentes en la aplicación del *fair use* al entrenamiento de modelos de GenAI con obras literarias protegidas. En ese sentido, debe señalarse que la doctrina y la jurisprudencia estadounidenses ofrecen dos enfoques diferenciados que no son necesariamente mutuamente excluyentes pero cuya tensión sólo será resuelta cuando los tribunales de apelación competentes emitan pronunciamiento sobre el fondo de la cuestión, pues, tenemos las contracaras de *Bartz v Anthropic* y *Kadrey et al. v. Meta Platforms*, donde un mismo asunto es tratado de diferentes formas, llegando a concluir que algunas veces sí se justifica por *fair use* la adquisición de obras protegidas por *copyright* en plataformas conocidas como *shadows libraries*, por otro lado, se continúan juicios por esta misma causa.

Dado que este fenómeno aún se encuentra en una etapa temprana de desarrollo jurídico y tecnológico, el presente TEG constituye una contribución analítica al debate contemporáneo sobre los límites del adiestramiento de sistemas con material protegido y la doctrina del *fair use*. Lejos de pretender obstruir la innovación tecnológica o negar el potencial de la GenAI, esta investigación busca fundamentar un marco de desarrollo armónico y paralelo entre las empresas desarrolladoras y los creadores. Esto resulta imperativo en un contexto donde los autores enfrentan crecientes dificultades para subsistir a partir de uno de los atributos más esencialmente humanos: la creatividad.

CAPÍTULO IV: DESARROLLO DE LA PROPUESTA

A continuación se enuncian algunos criterios relevantes para considerar los límites del aprendizaje de las GenAI por parte de las empresas desarrolladoras. No obstante, resulta necesario precisar que el presente TEG no aboga por un rechazo al progreso tecnológico ni por la limitación absoluta del desarrollo de sistemas de IA. Su propósito consiste en identificar parámetros jurídicos que permitan evaluar cuándo el uso de obras protegidas por copyright para el entrenamiento de modelos generativos puede afectar los derechos de los titulares, especialmente en aquellos supuestos en los que la producción creativa de autores, artistas o editores es utilizada como insumo para el desarrollo de productos comerciales explotados por grandes corporaciones tecnológicas.

Obtención de las obras utilizadas para el entrenamiento

Uno de los criterios más relevantes planteados a lo largo del presente TEG consiste en analizar la forma en que las obras son obtenidas para el entrenamiento de estas nuevas tecnologías.

“Una IA sin datos es un motor sin gasolina. Los datos se usan para entrenar a la IA y lograr que aprenda, por usar el argot del sector. El ML, no significa otra cosa que lograr que nuestro modelo aprenda a partir de ejemplos llamados observaciones. Si queremos crear y entrenar a una IA para que detecte gatitos en nuestra fotos, necesitamos mostrarle miles de fotos con gatitos en diferentes posiciones” (Fenollosa, 2024, p-92).

Es por esto que, para que el *output* sea lo más coherente posible, el sistema debe entrenarse con la mayor cantidad de información disponible, sin importar si esta se encuentra protegida por *copyright* o pertenece al dominio público.

Resulta fundamental recordar que la disponibilidad de contenido en internet no implica, por sí misma, que dicho material pertenezca al dominio público. Esta premisa adquiere especial relevancia frente al funcionamiento de los denominados *crawlers* -sistemas automatizados diseñados para recorrer la web y recopilar grandes volúmenes de información de diversa naturaleza-. En numerosos casos, las empresas desarrolladoras de GenAI se benefician de estos mecanismos de recopilación masiva de datos, aun cuando los titulares de derechos hayan intentado limitar el acceso mediante herramientas de protección como *robots.txt*, *NoAI meta tags* o *ai.txt*. Mientras siga existiendo un vacío normativo, persistirá una situación de considerable inseguridad jurídica en el entorno digital, donde prácticamente cualquier obra publicada en internet podría terminar siendo utilizada para el entrenamiento de los LLMs sin autorización expresa de los autores.

Es por esto que se debería analizar más este asunto y llegar a alguna prevención de los *crawlers* en la web, algún medio probatorio que evidencie cuándo una empresa ha adquirido información atravesando barreras de seguridad impuestas por los titulares de los derechos.

Con respecto a las *shadow libraries*, la piratería no es un tema moderno; esta lleva siglos y ha sido la némesis de la industria literaria desde sus inicios. Sin embargo, la lucha toma un giro distinto cuando esta piratería se vuelve un mecanismo utilizado para alimentar los LLMs de las empresas de GenAI. Resulta contradictorio rechazar estas prácticas y, al mismo tiempo, permitir las dentro de dichas corporaciones tecnológicas, especialmente cuando estas obtienen beneficios económicos crecientes sin reconocer a los titulares de *copyright*.

Reproducción de las obras durante el entrenamiento de los LLMs

El presente TEG tiene como objetivo analizar la infraestructura operativa de la GenAI. Esto obedece a que el usuario común suele percibir únicamente el resultado final (*output*) de estos sistemas, ignorando el complejo proceso de alimentación que ocurre bajo una profunda asimetría de información. En este entorno, corporaciones valoradas en miles de millones de dólares aprovechan el acervo de obras protegidas por *copyright* sin otorgar remuneración alguna a sus autores.

La estrategia de alimentar a estos sistemas con obras contemporáneas responde al objetivo de consolidar una ventaja competitiva en el mercado, garantizando que el usuario obtenga un resultado (*output*) o una experiencia lo más cercana posible a la cognición humana, además de mitigar cualquier asimetría de información que impida responder a un *prompt*. Es por ello que la aplicación de la noción de uso transformativo a los datos que nutren las GenAI constituye actualmente un terreno de alta litigiosidad e incertidumbre jurídica; la existencia de múltiples variables procesales ha impedido que los tribunales alcancen un consenso definitivo, especialmente respecto a los datos adquiridos mediante fuentes no autorizadas.

Dentro de los marcos que han ido configurando los litigios estadounidenses, los límites que el sistema jurídico está construyendo sobre el aprendizaje de las GenAI con obras protegidas resultan significativos y crecientes. Sin embargo, su aplicación es fragmentada, lenta y generalmente favorable a los actores con mayor capacidad para asumir riesgos litigiosos. La adecuación del marco normativo a la escala y velocidad del desafío tecnológico requiere una intervención legislativa que el *Copyright Act*, en su configuración actual, no puede proveer por sí solo.

A ello se añade que la analogía entre estos sistemas y un niño que aprende —utilizada en diversos casos— resulta reduccionista e insuficiente; no contribuye a esclarecer las conductas permitidas en futuros litigios. Aunado a ello, en el país de estudio no existe consenso sobre qué debe considerarse GenAI, ni qué normas resultan aplicables a estas tecnologías, evidenciándose un vacío jurídico que requiere especial atención.

Outputs y posibles desplazamiento en el mercado

A modo ilustrativo, basta con realizar un ejercicio práctico mediante Claude solicitándole, a través de una serie de prompts, el resumen de la obra *The Lost Night* de Andrea Bartz. El resultado obtenido podría evidenciar la capacidad del sistema para describir la trama principal, identificar giros argumentales, desarrollar el rol de los personajes dentro de la historia e incluso precisar determinados elementos narrativos y objetos utilizados por los protagonistas.

Sin embargo, este resultado no responde a una comprensión humana o consciente de la obra, sino a procesos estadísticos derivados del entrenamiento del modelo sobre grandes volúmenes de datos. Como se ha explicado anteriormente, los sistemas de GenAI operan mediante relaciones probabilísticas construidas a partir de información previamente incorporada durante el entrenamiento, dentro de la cual presumiblemente se encontraba la obra completa de Bartz.

Precisamente allí surge una de las principales controversias jurídicas del presente TEG: la utilización de obras protegidas por copyright sin autorización de sus titulares, incluyendo material obtenido mediante plataformas de piratería masiva o *shadow libraries*.

Como se ha analizado en los casos desarrollados en este TEG, el desplazamiento del mercado y la afectación en las ventas de las obras de los titulares del *copyright* no han sido demostrados empíricamente como un estándar general. Ello conlleva un análisis complejo respecto de cada una de las obras utilizadas durante el entrenamiento, lo que representa una ardua labor para jueces y litigantes.

Sin embargo, resulta relevante mencionar el experimento realizado en 1973 titulado *The Doctor Fox Lecture*, el cual consistía en:

“Un paradigma en la seducción académica; dicho análisis radicaba en contratar un actor para interpretar ser un profesional en cierta área y dictar clases: dando un paradigma de conferencia suficientemente impresionante, un grupo experimentado de educadores que participa en una nueva situación de aprendizaje puede sentirse satisfecho de haber aprendido a pesar del contenido irrelevante, contradictorio y sin sentido transmitido por el conferenciante.” (Donald, H.1973. p-630).

Por lo tanto, se intentaba demostrar que el carisma y la autoridad de quien enseña pueden influir en un auditorio completo y generar la sensación de aprendizaje aun cuando la información suministrada sea incorrecta.

Este estudio adquiere relevancia contemporánea al trasladar la figura del *Doctor Fox* al contexto de los sistemas de GenAI, donde el usuario, mediante un *prompt*, puede intentar sustituir procesos de estudio e investigación, asumiendo que el aparente grado de sofisticación del sistema garantiza la calidad del conocimiento producido.

Dicho esto, puede observarse cómo estos sistemas de GenAI son utilizados en numerosas ocasiones no como herramientas complementarias, sino como sustitutos completos de fuentes originales de información. Por otro lado, las empresas continúan alimentando sus tecnologías con inmensas cantidades de datos; entre dichos materiales se incluyen “libros y artículos, con el fin de aprender cómo se comunican mediante el lenguaje las personas” (ChatGPT. s/f).

El desplazamiento del mercado constituye un aspecto que deberá seguir analizándose en el futuro cercano. Aunque no resulta irrazonable considerar la existencia de una posible afectación real, las pruebas deberán ser suficientemente contundentes y empíricas para construir estándares aplicables a las futuras prácticas de las empresas tecnológicas.

Relevancia práctica para Venezuela

Dicho todo esto, se plantea la pregunta sobre qué puede aprender el sistema venezolano respecto a los avances de EE.UU con respecto al entrenamiento de los LLMs y la resolución de los casos más relevantes. A pesar de que en Venezuela no existe la figura del *fair use* y que hasta la fecha no se ha presentado ningún caso de interés sobre el tema, la Ley Sobre Derechos de Autor (LSDA) menciona sobre cuáles actos pueden realizarse sin autorización del titular, lo cuales no dependen de una apreciación judicial amplia, sino de los supuestos expresamente, estas limitaciones de los derechos de explotación se encuentran establecidos desde el artículo 43 al 50, entre los cuales las relevantes a fines de este TEG incluyen:

“Usos con fines científicos, fines educativos, la reproducción individual de una obra por bibliotecas o archivos que no tengan fines de lucro, la copia de obras de arte efectuada a los solos fines de un estudio, la reproducción de las noticias del día o de hechos diversos que tengan carácter de simples informaciones de prensa, publicados por ésta o por radiodifusión, siempre que no constituyan obras de ingenio en razón de la forma y sin perjuicio de los principios que rigen la competencia desleal” (LSDA. 1993. Art 43-49).

Dicho esto, la LSDA expone diversos casos específicos en los cuales la ley excluye determinados actos del ámbito de los derechos exclusivos del titular, permitiendo el uso de una obra protegida sin necesidad de obtener autorización previa por parte de los titulares.

El sistema venezolano no se limita exclusivamente a estos artículos, ya que al formar parte del Convenio de Berna para la Protección de las Obras Literarias y Artística, en su artículo 9.2 concede que cada legislación permita la autorización siempre y cuando la “reproducción no atente a la explotación normal de la obra ni cause un perjuicio injustificado a los intereses legítimos del autor” (Convenio de Berna. Art. 9.2), de la cual se entiende la prueba de los tres pasos, la cual “es el instrumento fundamental del derecho internacional de autor para examinar la legitimidad de las limitaciones nacionales en materia de derechos de autor”(Schonwetter. s/f. párr. 26).

La regla de los tres pasos consiste en permitir las limitaciones en aquellos casos que involucren “en ciertos casos especiales; que no entren en conflicto con la explotación normal de la obra; y que no perjudiquen de manera irrazonable los intereses legítimos del autor/titular de los derechos” (Schonwetter. s/f. párr. 5)

Esta regla de los tres pasos no constituye una excepción autónoma ni autoriza directamente el uso de obras protegidas. Su función consiste en servir como parámetro para evaluar la compatibilidad de las excepciones previstas en la legislación nacional con las obligaciones internacionales asumidas por los Estados contrayentes. En consecuencia, incluso cuando un uso se encuentre contemplado por la LSDA, la interpretación también debe realizarse de forma compatible con dicho estándar internacional.

Dicho esto, se puede considerar que en ausencia de una excepción específica para el TDM y/o técnicas como *crawling*, el marco jurídico venezolano presenta importantes desafíos para determinar la legalidad del entrenamiento de sistemas de GenAI utilizando obras protegidas por el derecho de autor; el sistema venezolano responde a una lógica distinta a EE.UU con su respectivo *fair use* casuístico.

En consecuencia, la legalidad del entrenamiento de los LLMs con obras protegidas por Derechos de Autor dependerá de qué pudiera interpretarse dentro de alguna excepción expresamente prevista en la LSDA y, además, de que dicha interpretación respete la regla de los tres pasos del Convenio de Berna. A menos que el Estado venezolano decida adaptarse a leyes diferentes como la AI Act de la Unión Europea o decida crear una legislación propia para esta nueva modernidad tecnológica, sin embargo, ejemplos a su alrededor tiene, solo queda ver qué camino será mejor recorrer.

CAPÍTULO V

RECOMENDACIONES Y CONCLUSIONES

El presente TEG tuvo como objetivo general Analizar las implicaciones legales del aprendizaje de las GenAI obras protegidas por *copyright* en el sistema de EE.UU, con el fin de delimitar el *fair use* y su posible aplicación.

El objetivo planteado surge de la necesidad de proteger a los titulares de derechos ante las nuevas empresas de tecnología que cada vez se arraigan más a la sociedad moderna, donde la alimentación de estos grandes LLMs resulta en una dilución de aquellas obras protegidas por *copyright*.

Actualmente no existen estándares en los EE.UU por lo cual resulta imprescindible mantenerse en constante revisión de la base de datos de Edward Lee titulada “*Chat GPT Is Eating the World*”, con el fin de tener un seguimiento más cercano con las futuras decisiones que rigen esta nueva revolución tecnológica.

Para los futuros análisis en materia de *copyright* y GenAI, resulta imperativo establecer una premisa fundamental de interdependencia: sin *copyright*, no hay obras; sin obras, no existe material para el entrenamiento de los LLMs; y sin dicho entrenamiento, la GenAI carece de sustento operativo.

La preservación de la creatividad humana es, en última instancia, condición de posibilidad de la propia tecnología examinada en el presente TEG. La pregunta que H. R. Giger plantea: ¿estamos perdiendo nuestra esencia, estamos siendo abrumados por la tecnología?; no admite una respuesta definitiva al cierre de este análisis. Admite, en cambio, una toma de posición: los límites del aprendizaje de la GenAI no son un obstáculo al progreso, sino la condición necesaria para que ese progreso conviva con la dignidad económica y moral de los titulares de *copyright*.

Los estándares de este trabajo fueron hechos basados en el sistema de EE.UU y la figura del *fair use*, sin embargo, como aquella célebre frase de Shakespeare “hay más cosas en el cielo y en la tierra, Horacio, de las que han sido soñadas en tu filosofía”; existe un mundo entero fuera del sistema anglosajón que también debe avanzar y decidir en base a sus leyes y libertades. Este trabajo sirve como ventana, para que podamos analizar cómo EE.UU lleva el auge de las GenAI y que cada sistema jurídico pueda retroalimentarse con el fin de lograr una armonía entre tecnología y humanos.

Un ejemplo de esto es el IA Act de la Unión Europea que aunque no regule de forma explícita el entrenamiento de los LLMs, sí establece “que uno de los deberes de los distribuidores de IA, sería elaborar y poner a disposición del público un resumen suficientemente detallado sobre el contenido utilizado para el entrenamiento del modelo de IA de propósito general, de acuerdo con una plantilla facilitada por la Oficina de IA” (AI Act, 2024, art. 3, numeral 53d). Esto en contraposición con la legislación norteamericana, establece una responsabilidad tácita hacia las empresas de suministrar todo tipo de detalle de cómo surgen sin LLMs, y así, permitir una mayor transparencia sobre estas.

Adicionalmente, la Unión Europea en su directiva 2019/790 establece que “los desarrolladores de IA deben determinar primero si existe una reserva adecuada antes de entrenar sus modelos de IA y, de ser así, deben adquirir una licencia del titular de los derechos de autor.” (Stober & Dornis, 2026. p-3). La reserva aquí mencionada hace mención a los métodos planteados durante el TEG como Robots.txt para impedir el acceso “fácil” mediante crawlings, que, en caso de ser así, la directiva exige un contrato de licencia obligatoria para el entrenamiento de los LLMs en caso de corresponder.

Dicho esto, podemos ver como existen otras visiones que pueden dar mayores aportes y un camino más ilustrativos, esto sería relevante para otro TEG o análisis correspondiente, los LLMs y su entrenamiento es realmente algo nuevo de lo cual no se ha llegado a una homogeneización sobre cómo deben actuar las grandes empresas, y, he aquí donde los especialistas en propiedad intelectual, deben salir a relucir con diferentes opiniones que logren una mejor claridad en este camino tecnológico nublado.

REFERENCIAS

Pope, A. (2024). NYT v. OpenAI: The Times's About-Face. Harvard Law Review.
<https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>

Naftulin, D. H., Ware, J. E., Jr., & Donnelly, F. A. (1973). The Doctor Fox lecture: A paradigm of educational seduction. *Journal of Medical Education*, 48(7), 630–635.

Carroll, M. W. (2019). Copyright and the progress of science: Why text and data mining is lawful. *UC Davis Law Review*, 53(2), 893–964.
https://digitalcommons.wcl.american.edu/facsch_lawrev/1203/

Martínez, J. L. (2026). El hilo y el signo: IA y la búsqueda de una ética de la profundidad. Substack. <https://substack.com/home/post/p-199192342>

All-in-One AI. (2026). ChatGPT user growth: From 1M to 800M in 3 years.
<https://all-in-one-ai.co/chatgpt-user-growth>

Anthropic. (s. f.). Company. <https://www.anthropic.com/company>

Baack, S. (2024). *A critical analysis of the largest source for generative AI training data: Common Crawl*. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery.
<https://doi.org/10.1145/3630106.3659033>

Barraza Macías, A. (2023, enero). *Metodología de la investigación cualitativa: Una perspectiva interpretativa*. Benessere. Centro de Intervención para el Bienestar Físico y Mental A.C.
<https://www.upd.edu.mx/PDF/Libros/MetodologiaInvestigacion.pdf>

Bartz v. Anthropic PBC, No. 3:24-cv-05417-WHA (N.D. Cal. 23 de junio de 2025).
<https://copyrightalliance.org/wp-content/uploads/2025/06/Bartz-v.-Anthropic-Order.pdf>

Bender, E. M., Gebru, T., McMillan-Major, A., y Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?. En Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610-623). Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>

Bobadilla, J. (2020). Machine learning y deep learning. Ediciones de la U. <https://books.google.com/books?id=iAAyEAAAQBAJ>

Carroll, M. W. (2020). Copyright and the progress of science: Why text and data mining is lawful (American University Washington College of Law Research Paper No. 2020-15). American University Washington College of Law.

https://digitalcommons.wcl.american.edu/facsch_lawrev/1203/

Electronic Information for Libraries (EIFL). (s.f.). Copyright and text and data mining.

<https://www.eifl.net/programme/copyright-and-libraries-programme/copyright-and-text-and-data-mining>

Etimologías de Chile. (s. f.). Plagiar. Diccionario etimológico castellano en línea. <http://etimologias.dechile.net/?plagiar>

Feist Publications, Inc. v. Rural Telephone Service Co., 499 U.S. 340 (1991).

Gardner, G. J., McLaughlin, S. R., y Asher, A. D. (2017). *Shadow libraries and you: Sci-Hub usage and the future of ILL* [Ponencia]. ACRL 2017, Baltimore, Maryland, Estados Unidos.

<https://alair.ala.org/server/api/core/bitstreams/d5643d28-2b76-4d25-933d-308ceb759752/content>

Hariani, K., y Hariani, A. (2011). Analyzing "originality" in copyright law: Transcending jurisdictional disparity. IDEA: The Intellectual Property Law Review, 51(3), 491-510.

https://ipmall.law.unh.edu/sites/default/files/hosted_resources/IDEA/idea-vol51-no3-krishna-hariani-anirudh-hariani.pdf

Hernández Sampieri, C.R. (1991) Metodología de la investigación

https://www.uv.mx/personal/cbustamante/files/2011/06/Metodologia-de-laInvestigaci%C3%83%C2%B3n_Sampieri.pdf

Copeland, B. J. (Ed.). (2004). The essential Turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life. Oxford University Press.

Izquierdo, H. A. (2022). Artificial intelligence and text and data mining: Future rules for libraries?

En J. Coates, V. Owen & S. Reilly (Eds.), Navigating copyright for libraries: Purpose and scope (pp. 497–540). De Gruyter Saur. <https://doi.org/10.1515/9783110732009-022>

Knowledge Rights 21. (2024, April 23). *KR21 principles on artificial intelligence, science and research*.

<https://www.knowledgerights21.org/news-story/kr21-principles-on-artificial-intelligence-science-and-research/>

Lee, E. (s. f.). ChatGPT Is Eating the World [Blog]. <https://chatgptiseatingtheworld.com/>

Lenovo. (s.f.). What is a petabyte? Recuperado el 20 de mayo de 2026, de

<https://www.lenovo.com/gb/en/glossary/petabytes/>

Li, L. (2021). Response to WIPO conversation on intellectual property and frontier technologies.

World Intellectual Property Organization (WIPO).

https://www.wipo.int/documents/d/frontier-technologies/docs-en-pdf-interventions-ind_li.pdf

Litman, J. (2018). *What we don't see when we see copyright as property*. *The Cambridge Law Journal*, 77(3), 536–558. <https://doi.org/10.1017/S0008197318000600>

Liu, E., Luo, E., Shan, S., Voelker, G. M., Zhao, B. Y., & Savage, S. (2025). Somesite I used to crawl: Awareness, agency and efficacy in protecting content creators from AI crawlers.

<https://doi.org/10.1145/3730567.3732913>

Matich, D. J. (2001). Redes neuronales: Conceptos básicos y aplicaciones. Universidad Tecnológica Nacional, Facultad Regional Rosario.

https://www.frro.utn.edu.ar/repositorio/catedras/quimica/5_anio/orientadora1/monograias/matich-redesneuronales.pdf

Michael D Murray.(2023). Generative AI Art: Copyright Infringement and Fair Use, 26 SMU SCI. & TECH. L. REV. 259.

Miró Llinares, F. (2007). El futuro de la propiedad intelectual desde su pasado. La historia de los derechos de autor y su porvenir ante la revolución de Internet. Revista de la Facultad de Ciencias Sociales y Jurídicas de Elche.

Morales, O. A. (2003). Fundamentos de la investigación documental y la monografía. Facultad de Odontología, Universidad de los Andes.

<https://webdelprofesor.ula.ve/odontologia/oscarula/publicaciones/articulo18.pdf>

National Science and Technology Council. Select Committee on Artificial Intelligence. (2019). The national artificial intelligence research and development strategic plan: 2019 update. Networking and Information Technology Research and Development Program. <https://www.nitrd.gov/pubs/National-AI-RD-Strategy-2019.pdf>

Rose, M. (1993). *Authors and owners: The invention of copyright*. Harvard University Press.

Rouhiainen, L. (2018). Inteligencia artificial: 101 cosas que debes saber hoy sobre nuestro futuro. Alienta Editorial.

https://proassets.planetadelibros.com/usuaris/libros_contenido/arxiu/40/39308_Inteligencia_artificial.pdf

Schur, N. S. (s. f.). Mathematical modelling and the Dr. Fox lecture. https://romanfrigg.org/wp-content/uploads/links/Dr_Fox_Lecture.pdf

Stober, S., & Dornis, T. W. (2026). Generative AI Training and Copyright Law. arXiv. <https://arxiv.org/abs/2502.15858>

The Editors of Encyclopaedia Britannica. (2026). Alan Turing. Encyclopaedia Britannica.
<https://www.britannica.com/biography/Alan-Turing>

Anguiano, J. M. (2023). Inteligencia artificial y copyright: Del dilema de Thaler a la doctrina «the right to read is the right to mine». Instituto de Derecho de Autor.
<https://institutoautor.org/el-instituto-autor-presenta-la-obra-inteligencia-artificial-y-copyright-del-dilema-de-thaler-a-la-doctrina-the-right-to-read-is-the-right-to-mine>

U.S. Copyright Office. (s. f.). U.S. Copyright Office fair use index.
<https://www.copyright.gov/fair-use/>

University of Sheffield. (s.f.). Clive Humby.
<https://www.sheffield.ac.uk/cs/people/academic-visitors/clive-humby>

Whittaker, M. (2025). La plantación y la computadora: Tres ensayos sobre IA, trabajo y esclavitud. Virus Editorial.

Patterson, L. R. (1997). Regents guide to understanding copyright and educational fair use. Journal of Intellectual Property Law, 5(1), 243–306.
<https://digitalcommons.law.uga.edu/jipl/vol5/iss1/6/>

Morales, O. A. (2003). Fundamentos de la investigación documental y la monografía. En N. Espinoza & Á. Rincón (Eds.), Manual para la elaboración y presentación de la monografía (pp. 1–14). Grupo Multidisciplinario de Investigación en Odontología, Facultad de Odontología, Universidad de Los Andes.
<https://hopelchen.tecnm.mx/principal/sylabus/fpdb/recursos/r125656.PDF>

Rao, S. S., & Estoesta, I. C. (2026, 13 de febrero). IP hot topic: Privilege preserved: OpenAI escapes forced disclosure of attorney communications in major copyright fight. Sterne, Kessler, Goldstein & Fox PLLC.
<https://www.sterneessler.com/news-insights/client-alerts/ip-hot-topic-privilege-preserved-openai-escapes-forced-disclosure-of-attorney-communications-in-major-copyright-fight>

Allyn, B. (2025, 26 de marzo). Judge allows “New York Times” copyright case against OpenAI to go forward. NPR.
<https://www.whqr.org/2025-03-26/judge-allows-new-york-times-copyright-case-against-openai-to-go-forward>

Chen, N. (2025, 9 de junio). Stolen stories or fair use? The New York Times v. OpenAI and the limits of machine learning. Columbia Undergraduate Law Review.
<https://www.culawreview.org/ddc-x-culr-1/nyt-v-openai-and-microsof>