International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Review Article)

The Legal Status of AI Training Data: A Cross-Jurisdictional Analysis of Copyright, Fair Use, and Text-and-Data Mining

Chaudhary Hamza Riaz *

Award-Winning Independent Legal Researcher, Department of Law, University of Law, Birmingham, United Kingdom.

International Journal of Science and Research Archive, 2026, 18(01), 935-949

Publication history: Received on 19 December 2025; revised on 24 January 2026; accepted on 27 January 2026

Article DOI: https://doi.org/10.30574/ijsra.2026.18.1.0166

Abstract

This study examines whether training artificial intelligence systems on large-scale datasets constitutes copyright infringement and how legal outcomes differ across the United Kingdom, European Union, and United States. Using a comparative doctrinal methodology, it analyses statutes, case law, and regulatory instruments alongside the technical stages of scraping, tokenization, and parameterization to identify where acts of reproduction arise. The findings show that AI training inherently involves copying, but the legality of that copying varies: the UK maintains the strictest regime with narrow exceptions, the EU permits training through structured TDM rules with opt-outs, and the US provides the broadest protection under fair use. This fragmented landscape creates significant uncertainty and compliance burdens for developers while offering limited clarity for creators seeking compensation or control. The study concludes that harmonized reforms, improved transparency, and clearer statutory definitions are essential to balance innovation with the rights and economic interests of creators.

Keywords: AI; Copyright; Training data; Text-and-data mining; Fair use; UK law; EU law; US law; Reproduction right; Machine learning; Digital regulation

1. Introduction

1.1. Contextual Background

The rapid expansion of generative artificial intelligence (AI) in recent years has profoundly reshaped the global digital landscape, with large language models (LLMs) such as GPT-4 and GPT-5, multimodal systems, and diffusion-based image generators emerging as transformative technological tools in both commercial and research environments. These models rely on the ingestion of massive quantities of data, typically drawn from publicly accessible sources through large-scale web scraping. Prominent datasets such as Books3, LAION-5B, and Common Crawl contain millions, and in some cases billions, of text, image, and audio files, many of which are protected by copyright. This data-driven expansion has generated intense legal and public controversy. Creators across multiple industries, including authors, musicians, visual artists, programmers, and news organizations have filed lawsuits or issued public objections, arguing that their works have been copied without consent and used to develop commercial systems that may undermine their economic interests. The debate also reflects wider socio-economic tensions. While AI developers assert that large-scale training is essential for innovation, creative industries express concern that automated systems capable of reproducing or mimicking human expression could displace human labor, depress income streams, or replicate copyrighted styles without attribution.

[*] Corresponding author: Chaudhary Hamza Riaz

1.2. The Central Legal Problem

At the core of this debate lies a fundamental legal question: whether the act of training AI systems constitutes copyright infringement. Training an AI model requires reproducing protected works, often in their entirety, onto digital servers before converting these works into numerical representations or parameters during the learning process. Developers argue that the resulting model does not retain expressive elements of the original works, as the training process produces abstract statistical relationships rather than human-readable copies. However, from a doctrinal perspective, the act of copying itself, regardless of the model's later outputs may fall within the scope of reproduction rights under UK, EU, and US copyright laws. The difficulty stems from the fact that copyright frameworks were developed long before machine learning existed, and legislatures did not anticipate technologies capable of ingesting millions of works simultaneously. As a result, courts and policymakers across jurisdictions disagree on whether training constitutes unlawful copying, a lawful form of text-and-data mining (TDM), or a permissible fair use. This fragmentation has produced significant legal uncertainty for developers, creators, and regulators.

1.3. Why This Question Matters

The question carries enormous practical significance. Modern datasets used for AI training contain billions of copyrighted works sourced from books, news archives, image libraries, and software repositories. At the same time, the commercial value of the generative AI industry now measured in the hundreds of billions of pounds, depends heavily on whether these training practices are legally permissible, particularly for models deployed across multiple jurisdictions. Because there is no international harmonization of copyright standards relating specifically to machine learning, companies must navigate an increasingly fragmented regulatory environment in which the legality of training may vary dramatically between the UK, EU, and US. This uncertainty affects not only developers but also creators, whose rights and revenue streams may be jeopardized without clear legal protection. The absence of a coherent international position has therefore created tensions between innovation, economic fairness, and fundamental rights, raising concerns about market distortion, unequal bargaining power, and the long-term sustainability of creative professions.

1.4. Research Aims

This research seeks to clarify the legal status of AI training data by undertaking a cross-jurisdictional analysis grounded in copyright doctrine. First, the study examines whether training AI models constitutes reproduction under existing UK, EU, and US copyright laws. Second, it evaluates the scope of relevant exceptions, including the American fair use doctrine, the EU's and UK's TDM exceptions, and related statutory provisions. Third, the study analyses major court cases and regulatory interventions that shape the debate, including decisions such as Authors Guild v Google, Infopaq, Pelham, and Warhol v Goldsmith. Finally, it proposes principles for developing a coherent and globally applicable legal framework that balances the interests of creators, industry, and the public.

2. Methodology

The paper employs doctrinal legal research methods to interpret statutory provisions, judicial decisions, and regulatory documents across the three selected jurisdictions. The analysis is predominantly comparative, identifying similarities, differences, and conflicts between the UK, EU, and US approaches to AI training and copyright. No empirical or technical experiments are undertaken; instead, the study focuses on analytical and interpretative examination of legal rules. The aim is to determine whether training falls within existing definitions of “copying,” “reproduction,” or “transformative use,” and to evaluate the adequacy of current exceptions. Ethical issues, non-copyright policy questions, and broader socio-technical concerns fall outside the explicit scope of this research, which is confined to the legal status of training under copyright.

2.1. Structure of the Paper

The paper proceeds in six main parts. Following this introduction, Section 5 explains the technical mechanics of AI training to clarify what constitutes “copying” in machine learning. Section 6 outlines foundational copyright concepts relevant to the analysis. Sections 7, 8, and 9 provide detailed examinations of the UK, EU, and US legal frameworks. Section 10 offers a comparative analysis across the three systems. Section 11 discusses creators' rights and concerns. Section 12 proposes policy reforms and future legal pathways. The paper concludes in Section 13 by summarizing key findings and offering final recommendations.

3. Technical Background How AI Training Actually Works

Artificial intelligence models such as large language models (LLMs) operate through a multi-stage training pipeline that transforms raw data into statistical parameters capable of producing human-like text. Understanding these technical steps is essential because each stage raises distinct legal questions regarding copyright, reproduction, temporary copies, and the status of model weights. This section outlines the main phases of AI training, data collection, tokenization, parameterization, and explains why these processes matter for legal analysis, particularly in the context of compliance with the right to a fair trial and copyright obligations.

3.1. Data Collection and Scraping

Most AI systems start with large-scale data collection, often through automated web crawlers that systematically scan, copy and store material from publicly accessible websites. This process, known as scraping, results in the mass ingestion of textual, visual and audio content. In the context of LLMs, training datasets may contain billions of words extracted from books, articles, blogs, court judgments, statutes, and other public or semi-public sources. Although developers often claim that only “public” sources are used, scraping frequently copies entire works, including copyrighted works into training corpora.

From a copyright perspective, scraping itself involves copying because raw data must be downloaded and stored before it can be processed by the model. Even if the scraping is temporary, the act of reproduction occurs at the moment the data is captured on a server. This raises questions about whether large-scale scraping constitutes lawful or unlawful reproduction, under “temporary copies” exceptions, and whether training a model amounts to “text and data mining” (TDM). UK law, unlike the EU, does not provide a comprehensive, open TDM exception for commercial AI developers, making this stage highly legally sensitive.

3.2. Tokenization and Pre-Processing

Once data is collected, the AI system converts text into numerical units using tokenization, a process that breaks language into tokens, sub words, characters, or word fragments. Tokenization requires the AI system to temporarily load and hold copies of the original text so that each segment can be transformed. During pre-processing, the system cleans, normalizes, and structures this tokenized data, discarding formatting, special characters, and unnecessary metadata.

The legal question is whether tokenization itself constitutes reproduction. Even though tokenization produces an abstract numerical representation, courts and scholars increasingly acknowledge that creating intermediate copies such as for indexing or transformation may still infringe copyright if done without permission. Under UK copyright law, reproductions “in any material form” include temporary digital copies made as part of a technological process, unless a specific exception applies. The transformation of text into tokens arguably requires reproducing the original work in RAM or local storage, thereby triggering the reproduction right. Thus, this technical stage has significant implications: AI developers cannot avoid infringement merely by converting works into tokens, because the copying occurs before abstraction happens.

3.3. Parameterization

After tokenization, the model undergoes training, during which it adjusts billions of internal parameters (weights) based on statistical relationships observed in the training data. The model does not store the original text; instead, it develops mathematical patterns governing how tokens relate to one another. This process is called parameterization.

- A central legal distinction emerges here:
- The model generalizes from data by adjusting statistical weights.
- It does not retain text verbatim, except in rare memorization cases involving highly repeated or unique content.

This raises the question: Are model weights copies of copyrighted works? Technically, weights contain numerical values representing correlations, not expressive content. Scholars generally argue that these are non-expressive abstractions, structurally analogous to “ideas” rather than “expression”. However, some evidence shows that models can unintentionally memorize training data, particularly shorter or unique texts. When such memorization leads to verifiable reproduction in output, this can constitute infringement.

Thus, the parameterization stage is where the law must distinguish between lawful generalization and unlawful memorization.

3.4. Importance for Law

Understanding the technical pipeline allows clearer analysis of the legal issues:

- Are model weights copies?

Most evidence suggests that model weights contain statistical correlations, not protected expression. On this basis, they are unlikely to be treated as "copies" However, this depends on factual questions about memorization.

- Does training create temporary or permanent reproductions?

Yes, Training involves multiple acts of reproduction: scraping, storing, tokenizing, caching, and loading data into memory. UK law only allows temporary, transient copies if they are integral to lawful use whether AI training qualifies is contested.

- Does AI output resembling training data imply infringement?

Not automatically Copyright infringement requires substantial similarity and causal connection Output resembling training data may result from:

- Legitimate generalization,
- coincidental similarity, or
- unlawful memorization.

Only the last category is likely to infringe, and determining which applies requires technical examination.

4. Foundational Copyright Concepts

Copyright law establishes a framework of exclusive rights that regulate how creative works may be used, copied, transformed, or disseminated Artificial intelligence training interacts with several of these rights, particularly reproduction, temporary copying, derivative works, and moral rights Because existing statutes were drafted long before machine learning existed, courts must determine how traditional copyright doctrines apply to inherently technical processes such as data scraping, tokenization, and parameterization This section outlines the core legal concepts relevant to assessing whether AI training infringes copyright, with a comparative overview of the United Kingdom, the European Union, and the United States.

4.1. The Right of Reproduction

The right of reproduction is the central copyright implicated by AI training In the United Kingdom, section 17 of the Copyright, Designs and Patents Act 1988 (CDPA) defines reproduction broadly as copying a work "in any material form", including storing it electronically. This means that even transient or machine-readable copies may constitute reproduction.

The EU takes a similar approach Article 2 of the InfoSec Directive grants authors the exclusive right to authorize or prohibit "direct or indirect, temporary or permanent reproduction by any means and in any form". This broad language captures the entire lifecycle of digital copying, including automated processes performed by AI systems.

In the United States, 17 U.S.C § 106 grants copyright owners the exclusive right to reproduce works "in copies or phonorecords", interpreted to include fixed digital copies. U.S courts have held that digital storage, even in RAM, may count as reproduction where the work is "sufficiently permanent" to be perceived or communicated.

Because AI training requires copying datasets into storage, loading them into memory, and transforming them through tokenization, these acts prima facie engage the reproduction right in all three jurisdictions. The key legal question is whether any statutory exception or doctrinal limitation applies to these copies.

4.2. Temporary and Incidental Copies

Temporary and incidental copies are central to determining the lawfulness of AI training because machine learning involves creating numerous short-lived copies as part of the training process.

EU Law: The Infomax judgment confirmed that even an eleven-word extract may constitute reproduction if it reflects the author's intellectual creation. More significantly, the Court held that temporary copies created by technical processes such as scanning and indexing, still fall within the scope of reproduction unless an exception applies The Meltwater case further held that browsing the internet creates temporary copies that may infringe unless justified under Article 5(1) exception for transient and incidental copies. In SAS Institute, the Court clarified that the functionality of computer programs and underlying ideas are not protected, but intermediate reproductions during reverse engineering may still fall under copyright protection.

United Kingdom (Post-Brexit): Post-Brexit, UK courts continue to interpret temporary copy exceptions in line with the EU approach, given that the CDPA incorporates the same language and remains influenced by EU jurisprudence. The requirement that a temporary copy be transient, incidental, and an integral part of a technical process narrows the scope of lawful copying Whether AI training qualifies as "transient and incidental" is contested, as training involves large-scale, deliberate copying rather than passive, automatic caching.

United States: The temporary copy doctrine in U.S law stems from cases such as MAI Systems v Peak, which held that loading software into RAM constitutes reproduction. However, later cases have softened this position AI developers may argue that training copies are functional and non-expressive, but this remains unresolved U.S courts analyzing AI cases (e.g., Thaler, Andersen v Stability AI) have not yet squarely addressed temporary copy doctrine in machine learning.

4.3. Derivative Works / Adaptation

The concept of derivative works determines whether transformations performed during AI training constitute the creation of new copyrighted works requiring authorization Under UK law, adaptations include translations, arrangements, and transformations that recast the original work in a new form. The EU's approach similarly requires that a derivative work expresses the author's intellectual creation.

In the United States, derivative works are defined in 17 U.S.C § 101 as works "based upon" pre-existing works that recast, transform, or adapt the original. The recent Warhol v Goldsmith decision emphasized that even transformative works must justify their use under a specific purpose and cannot rely solely on stylistic differences to qualify as fair use.

Whether AI training creates derivative works is debated Tokenization and parameterization do not create expressive output; they create statistical representations However, if training outputs internal representations that replicate or encode expressive features of a work, plaintiffs may argue that this constitutes a derivative work Warhol suggests U.S courts may apply a stricter approach toward what counts as a transformation, potentially affecting AI cases.

4.4. Moral Rights and Attribution

Although moral rights are not central to reproduction analysis, they are relevant to questions of authorship, integrity, and attribution The UK recognizes moral rights under sections 77-89 of the CDPA, including the right to be identified as the author and the right to object to derogatory treatment. The EU's approach derives from the Berne Convention and tends to provide stronger protection, especially in civil law jurisdictions.

AI systems that use a creator's work without attribution, or that generate distorted imitations, could implicate moral rights, particularly the right of integrity While U.S law provides limited moral rights protection, through the Visual Artists Rights Act (VARA), the doctrine may still influence courts' interpretation of harm caused by AI-generated works.

5. Jurisdictional Analysis United Kingdom

The United Kingdom provides one of the most detailed statutory frameworks governing reproduction, temporary copying, and text-and-data mining (TDM), yet it remains ill-equipped to address the unique challenges posed by large-scale AI training The Copyright, Designs and Patents Act 1988 (CDPA) remain the primary governing instrument, supplemented by a series of fair-dealing exceptions and the recent TDM exception inserted to implement the EU's Information Society Directive Despite this, AI developers face substantial uncertainty as to whether training constitutes infringement and whether any statutory defense applies This section examines the UK regime in detail.

5.1. Statutory Framework

The CDPA 1988 defines reproduction broadly as copying a work "in any material form," which includes storing a work electronically. This expansive language means that virtually all stages of AI training, from dataset ingestion to tokenization, fall prima facie within the scope of the reproduction right.

Several exceptions may be relevant:

- s.28A: permits transient and incidental copies forming an integral and essential part of a technological process.
- s.29(1): fair dealing for the purpose of non-commercial research or private study.
- s.29A: a specific exception for text and data mining for non-commercial research.
- s.30A: quotation exception, allowing limited reproductions for criticism or review.

However, all these exceptions are narrow, and none appear designed for industrial-scale AI training Even after Brexit, the UK has retained the wording and structure of these provisions, and courts still rely heavily on pre-Brexit EU jurisprudence when interpreting them.

5.2. Does AI Training Constitute Copying?

AI training requires storing protected works in memory, transforming them into tokens, and creating multiple intermediate copies during the computational process Based on statutory wording and case law, these acts amount to reproduction.

UK jurisprudence has consistently applied a broad definition of copying, particularly in digital contexts Courts have held that even transiently cached files or screen-display copies constitute reproduction when they embody sufficient elements of the original author's intellectual creation. Given this, training datasets loaded into GPU memory, intermediate token files, and pre-processed text all qualify as copies.

Because the CDPA's focus is on the act of reproduction itself, not on human perception, the fact that AI systems (not humans) read the works is irrelevant The key point is that the works are stored and processed in a material form Accordingly, UK law strongly supports the conclusion that AI training constitutes copying unless a specific exception applies.

5.3. The Text and Data Mining (TDM) Exception s.29A CDPA

Section 29A CDPA, introduced to transpose Article 5(3)(a) of the InfoSec Directive, permits making copies of works for the sole purpose of computational analysis, including text and data mining However, the exception is restricted to:

- Non-commercial research, and
- Users with lawful access to the material.

As a result, the exception does not cover commercial AI developers such as OpenAI, Google, Anthropic, or Stability AI The overwhelming majority of large-scale AI training is commercial in nature, meaning s.29A offers no defense.

Furthermore, the statutory language is outdated It presumes that computational analysis occurs in a controlled research environment, not in cloud-scale, self-learning AI models It also fails to address questions such as whether web-scraped data qualifies as "lawfully accessed," a key issue given ongoing litigation surrounding datasets like LAION-5B and Common Crawl.

5.4. The Temporary Copying Exception s.28A CDPA

Section 28A provides an exception for temporary and incidental copies that are:

- Transient,
- Ephemeral, and
- An integral and essential part of a technological process.

This exception was interpreted narrowly in cases such as Infomax and Meltwater, both of which continue to influence UK courts post-Brexit. AI training does not fit comfortably within this framework for several reasons:

- Training copies are not transient – datasets are stored for hours, days, or weeks.
- Copies are not incidental; they are deliberately created as part of the training objective.
- Training is not merely a technical process enabling lawful use, it is the core use itself.

Given these constraints, most scholars agree that s.28A does not excuse the copying involved in modern machine-learning workflows.

5.5. Government's 2022–23 TDM Reform Attempt

In 2022, the UK Government proposed a sweeping reform that would have permitted commercial TDM for any purpose, including AI training, without requiring rights-holder permission. The proposal triggered intense backlash from the creative industries, publishers, and the House of Lords Communications Committee, which described the plan as "reckless" and harmful to UK creators.

By February 2023, the Government formally withdrew the proposal This withdrawal maintains the status quo: commercial AI training remains unlicensed and unlawful. The episode also highlighted a deep policy divide whether the UK should become an AI-friendly jurisdiction by weakening copyright protections or maintain a creator-centric approach.

5.6. UK Position: Summary

The current UK legal position can be summarized as follows:

- AI training constitutes reproduction under the CDPA's broad definition.
- Commercial AI developers cannot rely on s.29A and are therefore unprotected.
- Temporary copy exception (s.28A) does not apply because training copies are neither transient nor incidental.

The Government's failed TDM reform attempt leaves significant legal uncertainty, discouraging innovation while also failing to protect creators in a coherent manner.

In its present form, UK copyright law is ill-adapted to large-scale machine learning Courts may eventually develop new doctrines, or Parliament may revisit TDM reform, but until then, AI companies face substantial legal risk when training models on copyrighted datasets within the UK.

6. Jurisdictional Analysis European Union

The European Union provides one of the most sophisticated and technologically conscious copyright frameworks relevant to AI training Unlike the UK and the US, the EU has explicitly addressed text and data mining (TDM) within its legislation, particularly through the 2019 Digital Single Market (DSM) Directive While the EU's underlying copyright principles remain grounded in the Information Society (InfoSec) Directive, more recent reforms demonstrate a deliberate attempt to accommodate automated analysis, large-scale computational research, and AI innovation Still, EU law preserves substantial authorial control through mechanisms such as opt-outs and strict conditions for lawful access This section evaluates the relevant directives, case law, and the emerging role of the EU AI Act in shaping copyright compliance obligations for generative models.

6.1. InfoSec Directive (2001/29/EC)

The starting point for EU copyright analysis remains the InfoSec Directive, which adopts an intentionally broad definition of reproduction Article 2 grants authors the exclusive right "to authorize or prohibit direct or indirect, temporary or permanent reproduction" of their works. This language has been interpreted expansively by the Court of Justice of the European Union (CJEU), which has confirmed that even small extracts or temporary digital copies can constitute reproduction where they express the author's intellectual creation.

For AI training, this means that every stage of dataset processing, scraping, tokenization, catching, embedding generation, constitutes reproduction within the meaning of Article 2 The Directive does not distinguish between copying for human consumption and copying for machine analysis; what matters is the act of fixation in material form As a result, AI training falls squarely within the scope of the reproduction right unless subject to an exception.

6.2. DSM Directive 2019: TDM Exceptions

The DSM Directive 2019 introduced two dedicated TDM exceptions that significantly reshape the legality of AI training within the EU:

- Article 3 TDM for Research Organizations and Cultural Heritage Institutions
- Article 3 provides a mandatory exception for TDM conducted by:
  - Research organizations, and
  - Cultural heritage institutions,

for the purposes of scientific research, provided they have lawful access to the works. Member States cannot override or opt out of this exception This creates a clear safe harbor for universities and public research labs training machine-learning models on copyrighted materials.

- Article 4 TDM for Any Purpose (Opt-Out by Rights Holders)
- Article 4 introduces a much broader exception allowing TDM for any purpose, including commercial AI development, unless rights-holders actively opt out. The opt-out must be expressed "in an appropriate manner," typically via machine-readable means such as metadata or website directives.

This provision is critically important: unlike the UK or US, the EU provides a pathway for commercial AI training to occur lawfully, so long as the mined works have not been opt-ed out by their creators.

6.3. The Opt-Out Principle

The EUs opt-out system is built on the assumption that creators should be able to retain control over whether their works are mined by automated systems.

However, several practical issues arise:

- Metadata Requirements: Many creators lack the technical knowledge or infrastructure to embed TDM-blocking Metadata into their works.
- Robots.txt Limitations: Robots.txt signals are often ignored by large-scale scrapers and do not necessarily bind downstream dataset creators.
- Enforcement Challenges: Once a dataset is created and redistributed, determining whether any item was opt-ed out is extremely difficult.

As a result, although Article 4 theoretically empowers authors, its real-world effectiveness is uneven, creating compliance challenges for AI companies operating across the EU.

6.4. Relevant Case Law

Several key CJEU decisions shape how AI training is assessed under EU copyright law:

- Infomax (C-5/08): Established that even short textual extracts can be protected and reaffirmed that temporary digital copies constitute reproduction.
- Pelham (C-476/17): Held that using even very brief sound samples may infringe unless the sample is unrecognizable, underscoring a strict approach to reproduction.
- Meltwater (C-360/13): Clarified the temporary-copy exception, ruling it applies only when copies are transient, incidental, and made for lawful use.

These decisions collectively reinforce that AI training constitutes reproduction, but temporary-copy exceptions are too narrow to excuse large-scale machine-learning processes Thus, compliance typically depends on the DSM TDM exceptions.

6.5. EU AI Act Implications

The EU AI Act, adopted in 2024, does not modify copyright law directly but introduces critical transparency and documentation obligations that significantly affect generative AI developers.

Key obligations include:

- Copyright-risk documentation: AI developers must document dataset sources and assess potential copyright conflicts in model training.
- Dataset transparency: Providers of general-purpose AI (GPAI) models must disclose summaries of training data and maintain records enabling traceability.
- Generative AI disclosure duties: Developers must ensure that outputs are marked or identifiable where necessary and disclose the use of copyrighted content in training when required.

While these provisions do not determine whether training is lawful, they increase accountability and reduce opacity, making it easier for rights-holders to enforce copyright claims.

6.6. EU Position: Summary

The EU offers the strongest regulatory clarity among major jurisdictions The InfoSoc Directive makes clear that AI training constitutes copying, but the DSM Directive provides explicit TDM exceptions that allow such copying to occur lawfully under defined conditions Article 3 protects academic research, while Article 4 enables commercial AI development subject to opt-outs The EU AI Act further enhances transparency and compliance obligations for generative models As a result, the EU stands out as the most structured and coherent legal environment for assessing the copyright implications of AI training.

7. Jurisdictional Analysis United States

The United States represents the most consequential jurisdiction for evaluating the legality of AI training, largely because the country's copyright system is uniquely shaped by the doctrine of fair use, an open-ended defence codified in 17 U.S.C §107. While Title 17 of the United States Code provides the baseline statutory framework, granting copyright holders exclusive rights including reproduction, distribution, adaptation, and performance under §106, the practical outcomes of litigation involving AI training hinge almost entirely on whether courts consider the ingestion of copyrighted works for model development to fall within fair use The US system does not contain a dedicated exception for text and data mining, unlike the UK or EU; instead, courts determine permissibility through a case-by-case assessment This creates both flexibility and unpredictability, allowing AI developers to rely on transformative-use reasoning while simultaneously exposing them to fact-specific litigation risks.

A central feature of US copyright law is that fair use acts as a balancing mechanism, enabling socially valuable uses of works that would otherwise infringe copyright The four-factor test, purpose and character of the use, nature of the copyrighted work, amount and substantiality of the portion used, and the effect upon the potential market is applied holistically. AI developers argue that training constitutes a highly transformative use because the works are not being consumed by humans but analyzed computationally to learn statistical patterns This reasoning parallels the judicial logic in previous cases that permitted large-scale copying for non-expressive analytic purposes However, recent jurisprudential developments, especially after the Supreme Court's decision in Warhol v Goldsmith, suggest that courts may no longer treat transformative purpose as determinative in the same manner as before, complicating the legal landscape for AI training.

7.1. Key Cases

The most significant precedent supporting the legality of AI training under US law remains Authors Guild v Google, commonly known as the Google Books case In this case, Google scanned millions of books without permission to create a searchable index and offer "snippet" views Both the Second Circuit and the Supreme Court (by declining certiorari) effectively endorsed the view that mass digitization for non-expressive, analytical purposes constituted fair use. The court emphasized that Google's use was transformative because it enabled search, research, and linguistic analysis rather than substituting for the original works This logic strongly supports the position that the creation of training datasets is permissible when the outputs do not reproduce expressive content in a way that competes with the source material.

However, this expansive reading of transformative use was narrowed substantially by the Supreme Court in Warhol v Goldsmith (2023), where the Court held that Andy Warhol's creation of a silkscreen image based on a photograph was not inherently transformative merely because it conveyed a different meaning. The Court shifted emphasis back to the commercial nature of the secondary use, suggesting that courts must carefully examine whether the purpose of the secondary use directly substitutes for or competes with the original market For AI, this means that courts may scrutinize whether model outputs compete with the economic markets of the works used in training such as images, writing styles, or music The Warhol decision therefore injects a level of uncertainty, even though its facts differ markedly from the non-expressive analysis involved in machine learning.

The emerging wave of litigation, GitHub Copilot, Stability AI, OpenAI, and other generative-AI lawsuits further illustrates the unsettled state of US law Plaintiffs have argued that AI outputs sometimes replicate training data and that the ingestion of datasets containing copyrighted works constitutes unauthorized reproduction. Defendants respond that training is analogous to the scanning in Google Books, where wholesale copying was permitted because the purpose was computational analysis rather than expressive use Since these cases remain pending, they represent one of the most consequential determinants of how US courts will interpret AI training under Title 17.

7.2. Analysis of AI Training Under the Fair Use Doctrine

Under the first factor, purpose and character of the use, AI training appears transformative because the works are converted into numerical tokens and weights, serving to develop generalized statistical models rather than substitute expressive content. However, commercial AI systems face heightened scrutiny after Warhol, as courts may consider whether output competes with creative markets The second factor, the nature of copyrighted work, typically disfavors fair use because training datasets often include highly creative works, but courts have historically given this factor limited weight The third factor, the amount used, presents a challenge because AI training requires copying entire works, but Google Books established that complete copying can still be fair when necessary for transformative analysis Finally, the fourth factor, the market effect, is increasingly contested Creators argue that generative AI tools erode markets for art, writing, and code, whereas developers contend that competition arises only from user prompts and outputs, not from the training process itself Overall, the fair-use assessment is mixed but still leans toward permissibility based on historical precedent.

7.3. US Position: Summary

The United States remains the most favorable jurisdiction for AI developers due to the flexibility and breadth of fair use While Google Books provides powerful support for the legality of large-scale data ingestion, Warhol and emerging litigation introduce new uncertainties, particularly concerning commercial substitution and market impact On balance, AI training is likely to be held lawful, but not with the degree of certainty seen in the EU's statutory TDM framework The US therefore offers the strongest potential protection for AI companies, but also the most high-stakes litigation environment, where courts hold broad discretion and outcomes may vary based on factual contexts.

8. Comparative Analysis

The legal treatment of AI training across the United Kingdom, European Union, and United States highlights significant divergences in statutory frameworks, exceptions, and judicial interpretations. First, the definition of reproduction varies considerably In the UK, the Copyright, Designs and Patents Act 1988 grants authors the exclusive right to control reproduction, and courts adopt a broad understanding of "copy," encompassing both temporary and permanent digital reproductions. Consequently, AI training likely constitutes infringement unless narrow exceptions apply. In the EU, the InfoSoc Directive provides a similarly broad definition of reproduction, but the subsequent DSM Directive introduces explicit exceptions for text-and-data mining, including an opt-out mechanism for rights holders. This dual system ensures that copying for research or commercial purposes is generally lawful if exceptions are correctly invoked, providing clearer legal guidance for developers. The United States, in contrast, relies on the flexible doctrine of fair use, where courts evaluate the purpose, nature, amount, and market effect of copying on a case-by-case basis. While this framework permits transformative AI training, outcomes are fact-specific and litigation-dependent, leading to strong protection in some contexts but uncertainty in others.

Second, the exception structures across jurisdictions further differentiate their approach The UK provides highly restrictive exceptions, primarily allowing non-commercial research TDM and limited temporary-copy defenses. The EU offers a structured exception regime: Articles 3 and 4 of the DSM Directive permits both academic and commercial TDM while granting rights holders opt-out control. The US relies on flexible fair use, giving courts broad discretion to balance the public interest in innovation against the author's rights, and often favoring transformative uses such as AI training.

Third, comparative risk assessment underscores the practical consequences for AI developers In the UK, the high likelihood of infringement and narrow statutory exceptions place developers at the highest legal risk, particularly for commercial applications. In the EU, risk is moderate, as compliance with TDM exceptions and careful attention to opt-outs allows developers to operate with relative certainty. The US presents the lowest risk in principle, given the historical protection for transformative uses under fair use; however, emerging litigation and the narrowing of "transformative use" in recent decisions inject potential volatility.

Table 1 Comparative overview of the legality of AI training under copyright law in the United Kingdom, European Union, and United States

Jurisdiction	Legality of AI Training	Exceptions	Conditions	Uncertainties	Notable Cases
UK	Likely infringement	Narrow TDM, temporary-copy exceptions	Non-commercial focus; transient copies	High – commercial AI excluded, statutory language outdated	Public Relations Consultants v NLA
EU	Generally lawful under TDM	Articles 3–4 DSM Directive; opt-out	Must comply with metadata/opt-out; research or commercial allowed	Moderate – enforcement and metadata issues	Infopaq, Pelham, Meltwater
US	Likely lawful under fair use	Flexible fair use doctrine	Transformative purpose; market effect assessment	Moderate – case-specific, Warhol v Goldsmith introduces nuance	Authors Guild v Google, Warhol v Goldsmith

Finally, the fragmented global landscape presents systemic challenges. Lack of harmonization among these major jurisdictions creates compliance burdens for developers operating internationally, as each legal regime demands different documentation, risk management, and adherence to exceptions. Moreover, inconsistent interpretations of reproduction, derivative works, and fair use mean that cross-border deployment of AI systems is fraught with legal uncertainty, potentially stifling innovation and increasing transaction costs. A harmonized, transparent, and globally coordinated framework would significantly reduce these burdens while respecting the rights of creators and the legitimate interests of AI developers.

9. Rights of Creators

The rise of AI-generated content raises significant questions regarding the rights and interests of creators whose works are included in training datasets. One of the central issues is compensation Creators across creative industries, authors, musicians, visual artists, and software developers demand payment when their works are used for AI training, arguing that large-scale copying can reduce licensing revenue and undermine their economic interests. While AI training does not involve human consumption of works, it may nonetheless create economic harm if the resulting models enable the generation of content that substitutes for the original, diminishes market demand, or reduces potential licensing fees. Different jurisdictions address these concerns inconsistently: the EU DSM Directive contemplates exceptions for research and commercial TDM without mandatory remuneration, while US fair use similarly allows transformative use without compensation, and the UK restricts commercial TDM under s 29A, limiting creators' recourse. Another critical dimension is consent Realistically, creators may find it challenging to opt out of datasets due to the opacity and scale of AI training. Data provenance is often unclear, and datasets like LAION-5B or Common Crawl include millions of works aggregated from the web without individual licensing, leaving authors with limited control. The opt-out mechanisms in EU law partially address this problem, but practical enforcement is uneven, especially for smaller creators who lack the resources to monitor and assert their rights.

Finally, moral rights remain a concern, particularly in jurisdictions that recognize attribution and integrity. Even when training is legally permitted, the output of generative AI may imitate the style of a particular creator without proper attribution, potentially infringing their right to be identified as the author or compromising the integrity of their work. While moral rights do not directly govern reproduction, they are essential to understanding the broader impact of AI on authorship, reputation, and the cultural ecosystem. Overall, ensuring that creators' economic and moral interests are respected requires careful legal design, including potential compensation models, transparent dataset registers, and enforceable opt-out mechanisms. Without such safeguards, the benefits of AI innovation risk coming at the expense of the very creators whose works fuel it.

10. Reform Options and Policy Recommendations

The growing use of AI in creative and industrial contexts necessitates a re-evaluation of copyright frameworks to balance innovation with creators' rights. One potential solution is the introduction of a collective licensing system, akin to those established in the music industry. Under such a model, creators could collectively license their works for AI training, enabling developers to access datasets legally while ensuring fair remuneration. This approach could streamline rights clearance, reduce litigation risk, and create a standardized market for training data, but feasibility depends on global cooperation and agreement on licensing rates. A complementary strategy involves establishing transparent dataset registers. Publicly accessible inventories of datasets would increase accountability by mandating disclosure of copyrighted sources used in AI training. Developers could certify compliance, authors could verify inclusion of their works, and regulators could monitor adherence to TDM or fair use exceptions Such registers would reduce opacity in dataset composition and enhance trust between creators and AI developers.

International harmonization represents another critical avenue Current fragmentation among the UK, EU, and US frameworks generates compliance burdens and legal uncertainty. Engagement through WIPO and adherence to the Berne Convention could provide a foundation for a minimum global standard for AI training data, harmonizing reproduction definitions, exceptions, and compensation mechanisms. A coordinated international framework would facilitate cross-border AI development while respecting authors' rights and market integrity. Revising TDM exceptions is equally important Expanding non-commercial TDM provisions and permitting commercial use subject to fair compensation would modernize statutory frameworks, particularly in jurisdictions like the UK where current exceptions are narrowly defined. Coupled with this, a clear legal definition of "AI training copy" would distinguish between functional reproductions, used solely for statistical learning and expressive reproductions that replicate creative content, providing legal certainty and reducing litigation risk.

Collectively, these reforms aim to create a balanced, transparent, and predictable legal environment for AI training. By combining licensing, transparency, international coordination, and statutory clarity, policymakers can support innovation while protecting creators' economic and moral rights, thereby fostering a sustainable ecosystem for both AI development and creative industries.

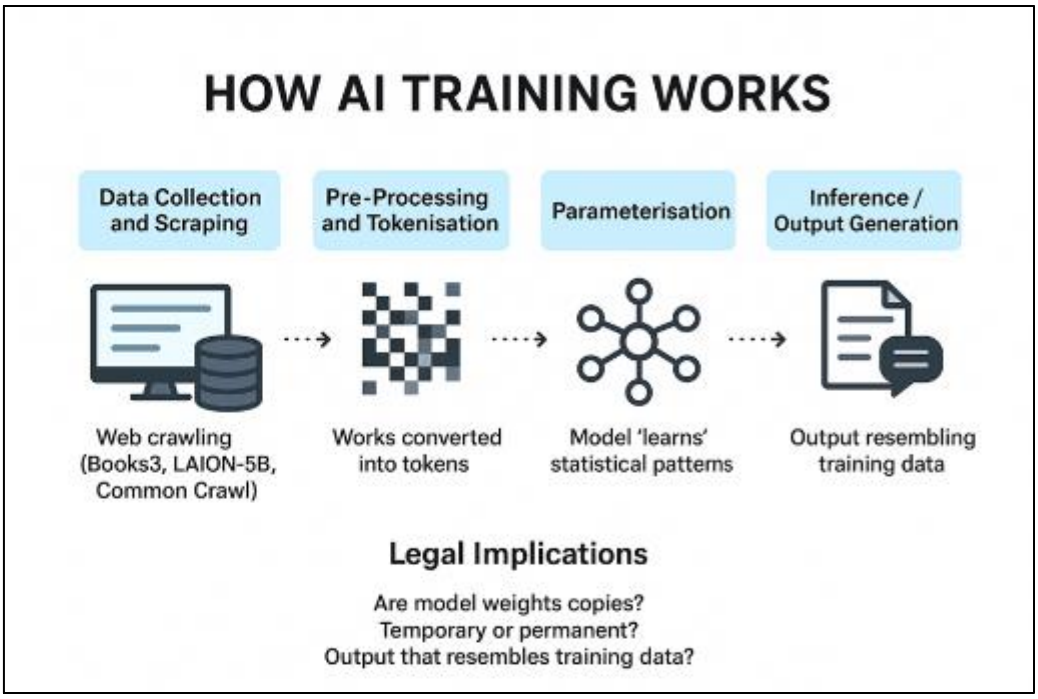
HOW AI TRAINING WORKS
Data Collection and Scraping
Pre-Processing and Tokenisation
Parameterisation
Inference / Output Generation
Web crawling (Books3, LAION-5B, Common Crawl)
Works converted into tokens
Model 'learns' statistical patterns
Output resembling training data
Legal Implications
Are model weights copies?
Temporary or permanent?
Output that resembles training data?

Figure 1 How AI Training Works

Table 2 UK–EU–US comparison of the legality of AI training

Jurisdiction	Legality of AI Training	Exceptions	Conditions	Uncertainties	Notable Cases
UK	Infringement	Narrow TDM, temporary-copy exceptions	Non-commercial focus; transient copies	High – commercial AI excluded, statutory language outdated	Public Relations Consultants v NLA
EU	Generally lawful under TDM	Articles 3–4 DSM Directive; opt-out	Must comply with metadata/opt-out; research or commercial allowed	Moderate – enforcement and metadata issues	Infopaq, Pelham, Meltwater
US	Likely lawful under fair use	Flexible fair use doctrine	Transformative purpose; market effect assessment	Moderate – case-specific, Warhol v Goldsmith introduces nuance	Authors Guild v Google, Warhol v Goldsmith

10.1. Definitions: Technical and Legal Terminology

- **AI Training Copy** – Any reproduction of original works (temporary or permanent) used in model training.
- **Tokenization** – Process of converting works into discrete units for machine learning.
- **Parameterization / Model Weights** – Internal representation of learned patterns derived from training data.
- **Text-and-Data Mining (TDM)** – Automated analysis of large datasets to extract patterns or information.
- **Fair Use** – US doctrine allowing limited use of copyrighted works without permission under specific conditions.
- **Temporary Copy / Incidental** Copy – Digital reproductions that exist only transiently for processing purposes.
- **Derivative Work / Adaptation** – New work based on pre-existing copyrighted material.
- **Opt-Out Mechanism** – Right for creators to prevent their works from being mined under TDM exceptions.

11. Conclusion

The analysis presented throughout this paper demonstrates that AI training inherently involves copying protected works, whether in the form of tokenization, parameterization, or temporary reproductions. While the technological process transforms the works into statistical representations rather than direct human-readable copies, the law in different jurisdictions treats these reproductions divergently, meaning that the legality of AI training is entirely jurisdiction-dependent. Examining the UK, EU, and US frameworks reveals distinct approaches. The UK maintains a restrictive framework, with narrow exceptions under s 29A CDPA and limited recognition of temporary-copy defenses, exposing AI developers to high legal risk. The EU provides a structured system, with DSM Directive TDM exceptions and opt-out mechanisms offering clarity and moderate legal certainty. The US relies on flexible fair use, often permitting transformative AI training, but outcomes remain fact-specific and contingent on judicial interpretation. These differences underscore the challenges of cross-border AI deployment and the need for harmonization.

This study contributes to scholarship by clarifying doctrinal uncertainties, offering a cross-jurisdictional comparison, and proposing policy solutions to reconcile creators' rights with AI innovation. By synthesizing statutory provisions, case law, and technical realities, it provides a framework for understanding how AI training interacts with copyright law and the practical implications for developers, creators, and policymakers. Finally, the paper argues that without reform, global AI development faces significant legal instability. A harmonized, transparent, and rights-balanced system is essential to ensure that creators' economic and moral rights are respected while enabling innovation in AI. Implementing collective licensing, transparent dataset registers, international coordination, and a clear legal definition of AI training copies can achieve this balance and provide the certainty needed for sustainable AI growth.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] Guadamuz A. Artificial intelligence and copyright. *European Intellectual Property Review*. 2017 Jan;39(1):12–18.

[2] Gervais D. The machine as author. *Iowa Law Review*. 2020 Apr;105(5):2053–2106.

[3] Ginsburg JC, Budiardjo LA. Authors, and machines: A comparative copyright analysis of human and artificial creativity. *Columbia Journal of Law & the Arts*. 2023 Winter;45(1):1–62.

[4] Leval PN. Toward a fair use standard. *Harvard Law Review*. 1990 Mar;103(5):1105–1136.

[5] Samuelson P. Why model weights are not copies under copyright law. *Journal of Intellectual Property Law & Practice*. 2023 Jun;18(6):421–433.

[6] Goldstein P, Hugenholtz B. *International Copyright: Principles, Law, and Practice*. 3rd ed. Oxford: Oxford University Press; 2019.

[7] Ricketson S, Ginsburg JC. *International Copyright and Neighbouring Rights: The Berne Convention and Beyond*. 2nd ed. Oxford: Oxford University Press; 2022.

[8] Giblin R, Doctorow C. *Chokepoint Capitalism: How Big Tech and Big Content Captured Creative Labor Markets and How We'll Win Them Back*. Melbourne: Scribe Publications; 2022.

[9] Patterson L, Lindberg S. *The Nature of Copyright: A Law of Users' Rights*. Athens (GA): University of Georgia Press; 1991.

[10] Elkin-Koren N, Salganik MJ. AI, data, and copyright: Reframing the boundaries of authorship and ownership. *Columbia Journal of Law & the Arts*. 2021 Fall;44(4):473–520.

[11] Halevy A, Norvig P, Pereira F. The unreasonable effectiveness of data. *IEEE Intelligent Systems*. 2009 Mar–Apr;24(2):8–12.

[12] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017 Dec; 30:5998–6008.

[13] Sennrich R, Haddow B, Birch A. Neural machine translation with subword units. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. 2016 Aug;1:1715–1725.

[14] Carlini N, Tramer F, Wallace E, Jagielski M, Herbert-Voss A, Lee K, et al. Extracting training data from large language models. *Proceedings of the 30th USENIX Security Symposium*. 2021 Aug;1:2633–2650.

[15] Biderman S, Schoelkopf H, Gao Q, et al. Datasets for large language model training: Transparency, scale, and governance. *NeurIPS Datasets and Benchmarks Proceedings*. 2022 Dec;1:1–22.

[16] Casey BJ, Lemley MA. Fair learning. *Texas Law Review*. 2021 Jun;99(4):743–785.

[17] Authors Guild v Google, Inc. 804 F.3d 202 (United States Court of Appeals for the Second Circuit, 2015).

[18] Andy Warhol Foundation for the Visual Arts, Inc v Goldsmith. 143 S. Ct. 1258 (United States Supreme Court, 2023).

[19] Infopaq International A/S v Danske Dagblades Forening. Case C-5/08 [2009] ECR I-6569 (Court of Justice of the European Union).

[20] Pelham GmbH v Hütter. Case C-476/17 EU:C:2019:624 (Court of Justice of the European Union).

[21] Public Relations Consultants Association Ltd v Newspaper Licensing Agency Ltd. Case C-360/13 EU:C:2014:1195 (Court of Justice of the European Union).

[22] SAS Institute Inc v World Programming Ltd. Case C-406/10 EU:C:2012:259 (Court of Justice of the European Union).

[23] MAI Systems Corp v Peak Computer, Inc. 991 F.2d 511 (United States Court of Appeals for the Ninth Circuit, 1993).

[24] Andersen v Stability AI Ltd. United States District Court for the Northern District of California; complaint filed 2023.

[25] Doe v GitHub, Inc. United States District Court for the Northern District of California; complaint filed 2022.

[26] Silverman v OpenAI, Inc. United States District Court for the Northern District of California; complaint filed 2023.

[27] Berne Convention for the Protection of Literary and Artistic Works. Paris Act, adopted 1886, as amended.

[28] Copyright, Designs and Patents Act 1988 (United Kingdom).

[29] Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society (InfoSoc Directive).

[30] Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market (DSM Directive).

[31] Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act).

[32] United States Code. Title 17—Copyright Law, §§101, 106, 107.

[33] House of Lords Communications and Digital Committee. *Artificial Intelligence and the Creative Industries*. London: UK Parliament; 2023.

[34] UK Intellectual Property Office. *Text and Data Mining: Legal Framework*. London: UK IPO; 2023.

[35] UK Intellectual Property Office. Consultation on Artificial Intelligence and Intellectual Property. London: UK IPO; 2022.

[36] World Intellectual Property Organization. Copyright and Artificial Intelligence Training Data. Geneva: WIPO; 2025.

[37] Common Crawl Foundation. Data repository overview. 2025. Available from: https://commoncrawl.org/ (accessed 1 December 2025).

[38] LAION. LAION-5B dataset documentation. 2025. Available from: https://laion.ai/ (accessed 1 December 2025).