IPR ISSUES IN MACHINE LEARNING DATASETS, CHALLENGES, AND EMERGING APPROACHES: A COMPARATIVE ANALYSIS BETWEEN THE US, THE EU, AND INDIA

Stephin Sinu Oommen, LLM (Intellectual Property and Trade Law), School of Law, CHRIST (Deemed to be University), Bengaluru[1]

ABSTRACT

With advances in machine learning (ML)-related technologies advancing rapidly, it creates complex intellectual property (IP) dilemmas about the ownership, use, and legal protection of the datasets used to train artificial intelligence (AI) systems. This paper addresses the intricate Copyright and database protection issues concerning machine learning datasets and examines issues including data scraping, copyrightability, and licensing regimes. The legal treatment of datasets varies from country to country internationally, leading to inconsistent layering of legal protections that collectively create barriers to providing appropriate incentives to innovate and protect IP rights. The paper compares how existing legal regimes apply to ML datasets in three large jurisdictions: the United States, the European Union, and India. Using selected legal cases from these jurisdictions, as well as EU policy instruments (Database and CDSM), India's Copyright Act of 1957, and US fair use doctrine, in particular, as it relates to text and data mining (TDM) practices, the paper introduces emerging legal practices, such as open datasets, ethical data sourcing, and potential avenues for creating more harmonized international legal harmonization.

The paper highlights the current legal gap, represented by the recent landmark decisions of ANI V Open AI in India, Authors Guild V Open AI in the USA, and Kneschke V LAION in Germany. It provides guidance on a smooth legal framework to clarify dataset ownership, usage rights, and licensing obligations with machine learning.

Keywords: Intellectual Property Rights, Artificial Intelligence Regulation, Machine Learning Datasets, Copyright Law

[1] Student, LLM (Intellectual Property and Trade Law), School of Law, CHRIST (Deemed to be University), Bengaluru

1. Introduction

The exponential growth of artificial intelligence and machine learning technologies has fundamentally transformed how data is collected, processed, and utilized for technological innovation.[2] The change is embedded in a complicated universe of intellectual property rights issues, and these challenges are testing the old legal systems to deal with pre-digital age content creation and distribution. Machine learning models, massive language models, and generative AI software need enormous amounts of training data, typically billions or millions of copyrighted pieces of work, proprietary databases, and personal data.[3] The change is a tangled web of intellectual property rights concerns that undermine established legal models for pre-digital age content creation and dissemination.[4]

Machine learning systems, massive language models, and generative AI. This sheer volume of data use has brought about what researchers call a "**collision between innovation and intellectual property rights**" that traditional legal systems are not well-suited to respond to.[5]

The body of law that regulates machine learning datasets is a patchwork of methods across various jurisdictions and divergent priorities to incentivize innovation as well as protect rights holders. In the US, courts have started tackling whether applying copyrighted works to train AI is a fair use under Section 107 of the Copyright Act[6], which requires huge training data, frequently consisting of millions or billions of copyrighted content, proprietary databases, and personal data.

The European Union has implemented specific text and data mining exceptions through the Copyright in the Digital Single Market Directive while maintaining robust database protection regimes.[7] India, meanwhile, operates under a copyright framework established in 1957 that

[2] A.B. Rashid, AI revolutionizing industries worldwide: A comprehensive review, 2 J. Artif. Intell. Ind. 100233 (2024), https://www.sciencedirect.com/science/article/pii/S2773207X24001386.
[3] M. Madanchian, The impact of artificial intelligence on research efficiency, 6 J. Innov. Data Sci. 8205 (2025), https://www.sciencedirect.com/science/article/pii/S2590123025008205.
[4] Poddar A, Rao SR. Evolving intellectual property landscape for AI-driven innovations in the biomedical sector: opportunities in stable IP regime for shared success. Front Artif Intell. 2024 Sep 17;7:1372161. doi: 10.3389/frai.2024.1372161. PMID: 39355146; PMCID: PMC11442499.
[5] Qin, Y., Xu, Z., Wang, X. et al. Artificial Intelligence and Economic Development: An Evolutionary Investigation and Systematic Review. J Knowl Econ 15, 1736–1770 (2024). https://doi.org/10.1007/s13132-023-01183-2
[6] 17 U.S.C. § 107 (2025) (fair use)
[7] Christophe Geiger, Giancarlo Frosio & Oleksandr Bulayenko, The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market – Legal Aspects, Policy Dep't for Citizens' Rts. & Const. Affs., Directorate-Gen. for Internal Policies, Eur. Parl. PE 604.941 (2018),

lacks explicit provisions for AI-related activities, leading to significant legal uncertainty as the country emerges as a central AI development hub.[8]

The stakes of resolving these legal uncertainties extend far beyond academic interest.[9] The AI industry has reached what some analysts describe as an inflection point where legal clarity around training data rights could determine the competitive landscape for the next generation of AI technologies.[10] At the same time, creators and publishers face serious existential questions regarding the value and protections of their creative works at a time when AI systems are able to reproduce, alter, or otherwise compete with human-created works.[11]

1.1 Research Questions

This comparative analysis addresses several critical research questions:

How do the legal frameworks in the United States, the European Union, and India address intellectual property rights in machine learning datasets, and what are the key differences in their approaches?

To what extent do existing copyright exceptions, fair use doctrines, and database rights adequately protect AI developers and rights holders?

What role do recent judicial decisions play in shaping the legal landscape for AI training data, and how do courts balance innovation interests with intellectual property protection?

How effective are emerging licensing models and regulatory approaches in addressing the unique challenges posed by large-scale data mining for AI purposes?

1.2 Research Objectives

[8] Gaurav Gupta & Nancy Roy, Text and Data Mining Vs. India's Digital Personal Data Protection Act, 2023: A Critical Study of the Legal Gap and Its Implications for AI Governance, Bridge Counsels Blog (Sept. 2025), https://bridgecounsels.com/text-and-data-mining-vs-indias-digital-personal-data-protection-act-2023-a-critical-study-of-the-legal-gap-and-its-implications-for-ai-governance/.

[9] C. Campbell, The AI Intelligence Playbook: Decoding GenAI Capabilities, 73 Bus. Horiz. 101602 (2025), https://www.sciencedirect.com/science/article/pii/S0007681325001405

[10] AI Consulting Director, What the Past Two Years of GenAI Mean for Legal's Future, DISCO Blog (2025), https://www.csdisco.com/blog/what-the-past-two-years-of-genai-mean-for-legals-future/.

[11] Michael Livermore & Daniel Rockmore, AI-Powered Lawyering: AI Reasoning Models, Retrieval Augmented Generation, and the Future of Legal Practice, (Mar. 2, 2025), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5162111.

The primary objectives of this research are to:

1. Offer an in-depth comparative analysis of intellectual property regimes covering machine learning datasets from three top jurisdictions.

2. Analyze the sufficiency of current legal principles in dealing with new challenges arising from AI development.

3. Examine current case law and regulatory changes to determine emerging trends and precedents.

4. Examine the effect and importance of the licensing and regulatory data protection approach.

1.3 Research Methodology

This paper employs a doctrinal legal method that analyses statutory provisions, case law, regulatory guidance, and academic commentary in the three jurisdictions. It evaluates recent developments in this area as well as ongoing lawsuits, including Authors Guild v. OpenAI in the US and ANI v. OpenAI in India. The study refers to primary legal sources, policy documents, and expert conversations in order to understand the evolving legal landscape.

2. Literature Review

2.1 Survey of Existing Literature

Artificial intelligence and intellectual property law have drawn significant academic interest in recent times. Initial academic research centered mainly on the patenting of AI inventions and the copyrightability of the content generated through AI. Nevertheless, as Prinsley et al. point out, the issue of the rights to training data has become the most critically relevant issue to the AI industry.[12]

Casale's examination of algorithmic transparency in public decision-making is pivotal in unpacking the "black box problem" that makes classic accounts of copyright infringement in

[12] Franscesco Schwab et al., AI Training and Copyright: Should Intellectual Property Law Allow Machines to Learn?, 9 Bioethica 45 (2024); Jessica Prinsley et al., Intellectual Property Rights and Training Data in AI, 18 Int'l J. Law & Tech. 223 (2024); see also Malavika Raghavan, Consent in the Age of AI: India's Missed Opportunities in the DPDPA, 21 Indian J.L. & Tech. 33 (2024).

AI cases more complicated. Her use of the Carltona principle to address algorithmic delegation offers a new approach to determining how legal accountability should be assigned when AI is used to process copyrighted content.[13]

Quang's thorough analysis of AI training and US copyright law articulates a persuasive case for statutory safe harbors for non-expressive uses of copyrighted works. Her distinction between expressive and non-expressive uses has set the stage for later judicial decisions and policy discussion.[14] Similarly, Shope's work on ethical requirements for the documentation of AI systems creates necessary guidance on transparency obligations that demystify some uncertainty regarding copyright.[15]

Kaminski and Urban's explanation about "right to contest AI" choices reflects a growing recognition that procedural safeguards are necessary to maintain both individual rights and group interests in the development of AI in the contemporary world. [16]

2.2 Unaddressed issues in the Literature

Despite great intellectual focus, a number of important gaps in the literature exist.

First, comparative analysis between multiple legal systems is still underdeveloped, with the bulk of the literature addressing individual jurisdictions.

Second, the fast speed of technological advancements implies that much of the existing work dates back to pre-current-generation AI systems and their record-breaking data commands.

Thirdly, the intersection of data protection and copyright laws is rarely factored into AI use, particularly where personal data is being utilized as training data.

3. Comparative Legal Frameworks

3.1 United States: Fair Use and Market-Based Solutions

[13] Elena Casale, Around the Black Box: Applying the Carltona Principle to Challenge Machine Learning Algorithms in Public Sector Decision-Making, 45 Oxford J. Leg. Stud. 727 (2024).
[14] Taylor Shope, Ethical Documentation and Transparency Obligations in AI Systems, 42 J. Info. L. & Tech. 75 (2025).
[15] Julie Quang, Does Training AI Violate Copyright Law?, 36 Berkeley Tech. L.J. 123 (2024).
[16] Margot E. Kaminski & Jennifer M. Urban, The Right to Contest AI, 121 Colum. L. Rev. 1957 (2021), https://scholar.law.colorado.edu/faculty-articles/1393/.

The United States mainly uses the doctrine of fair use contained in Section 107 of the Copyright Act to handle AI training data issues.

The four-factor fair use test examines:

(1) the purpose and character of the use,

(2) the nature of the copyrighted work,

(3) the amount used, and

(4) the effect on the market for the original work. [17]

Recent judicial decisions have shown an increasing willingness to find that AI training activities constitute fair use. In Bartz v. Anthropic (2025)[18], Judge Alsup found the use of legally obtained data for training LLMs to be "exceedingly transformative" because there was intent to identify statistical relationships rather than reproduce expressive content. The court in this case distinguished between legitimate digitization for training purposes and pirated materials, finding only the latter problematic under fair use analysis.[19]

The U.S. Copyright Office's 2025 report on AI training provides the most comprehensive official guidance to date. The Office came to the conclusion that while AI training may constitute prima facie infringement of the reproduction right, fair use analysis should focus on transformativeness and market harm because this also affects the market in some way. The Office rejected arguments that AI training is inherently transformative, emphasizing that the study must consider the specific use case and potential market substitution.[20]

3.2 Structured Exceptions and Database Rights in the European Union

The European Union has developed the most extensive legislative framework for addressing the training data of AI in the Copyright in the Digital Single Market Directive and the Database

[17] Copyright & Fair Use, Measuring Fair Use: The Four Factors, Stanford Univ. Libraries, https://fairuse.stanford.edu/overview/fair-use/four-factors/
[18] Bartz v. Anthropic PBC, No. 3:24-cv-05417 (N.D. Cal. June 23, 2025).
[19] Copyright and Fair Use: A Guide for the Harvard Community, Office of the Gen. Counsel, Harvard Univ., https://ogc.harvard.edu/pages/copyright-and-fair-use
[20] Copyright, AI Training, and Innovation, R Street Inst. (2025), https://www.rstreet.org/commentary/copyright-ai-training-and-innovation/.

Directive. Articles 3 and 4 of the CDSM Directive[21] create particular exceptions for text and data mining, where Article 3 deals with non-commercial research and Article 4 allows general TDM subject to opt-out provisions.

The seminal German ruling in Kneschke v. LAION (2024)[22] gave these provisions' first meaningful judicial explanation. The Hamburg Regional Court ruled that downloading and adding copyrighted photos to AI training datasets qualifies as part of the Article 3 TDM[23] exception for research institutions. In a notable ruling, the court dismissed such arguments that the possibilities of exploiting resulting AI models commercially exclude their use from the research exception protection.

The sui generis database right of the EU, under the Database Directive, offers further protection to significant investments in acquiring, checking, or presenting data. Recent scholarly commentary indicates that data produced by machines are outside such protection due to their nature as data creation and not acquisition. This makes a big difference for AI-created training sets.[24]

3.3 Evolving Framework and the Current Challenges in India

A major hurdle in the growth of AI development is the lack of targeted treatment of machine learning activities in the Indian copyright framework. This is due to the Copyright Act of 1957, Section 52,[25] which provides limited exceptions for fair dealing in the fields of news reporting, private use, and criticism. These exceptions do not offer the wide balancing of conflicting interests that the fair use doctrine does in the US.

The ongoing case of *ANI v. OpenAI*[26] before the Delhi High Court represents India's first significant test of how existing copyright law applies to AI training. ANI accuses OpenAI of

[21] Articles 3–4, of the European Parliament and of the Council of May 17, 2019, on Copyright and Related Rights in the Digital Single Market, 2019 O.J. (L 130) 92, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0790.
[22] Kneschke v. LAION eV, Higher Regional Court of Munich, Case No. 29 U 2876/23 (2024).
[23] Directive 2019/790, art. 3, of the European Parliament and of the Council of May 17, 2019, on Copyright and Related Rights in the Digital Single Market, 2019 O.J. (L 130) 92, https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32019L0790.
[24] Deepu Jacob Thomas & Prasan Dhar, Of Square Pegs and Round Holes: Towards a New Paradigm of Database Protection, 4 Indian J.L. & Tech. 35 (2008), https://docs.manupatra.in/newsline/articles/Upload/9AFF74BA-F91F-4F0E-ACE9-20CA86CBF941.pdf.
[25] Copyright Act, 1957, § 52, No. 14 of 1957 (India)
[26] Asian News International v. OpenAI Inc., W.P. (C) 123/2024, Delhi High Court (2024-2025).

copyright infringement by employing news content to train ChatGPT without permission. OpenAI's case is largely based on jurisdictional grounds, asserting that outside India, training exercises are beyond the scope of Indian copyright law.

The Digital Personal Data Protection Act 2023 adds new complexity as it regulates the processing of personal data employed in AI training. The Act does not require consent for publicly available data, but it places obligations on data fiduciaries that can impact the practice of AI development.[27]

In response to legal ambiguity, the government of India, in 2025, set up an expert committee to analyze whether the Copyright Act needs to be amended to handle AI-associated activities. The process of reviewing speaks volumes about the increasingly accepted awareness that the existing legal framework of India is insufficient for it as an AI development center.[28]

4. Analytical Framework: Text and Data Mining

4.1 Conceptual Framework

Text and data mining represent a fundamental shift in how copyrighted materials are utilized, from human consumption to automated analysis for pattern recognition and model training.[29] The European Union defines TDM as "*any automated analytical technique aimed at analysing text and data in digital form to generate information which includes but is not limited to patterns, trends and correlations*," and provides the most comprehensive statutory framework.[30]

The primary focus of legal interpretation in several jurisdictions is now the distinction between expressive and non-expressive uses of copyrighted works.[31] Supporters of broad TDM rights contend that removing statistical patterns from content protected by Copyright does not violate the core values that copyrighted legislation is intended to protect. Researchers assert that

[27] Raktima Roy & Gabriela Zanfir-Fortuna, The Digital Personal Data Protection Act of India, Explained, Future of Privacy Forum (2023), https://fpf.org/blog/the-digital-personal-data-protection-act-of-india-explained/.
[28] TrustArc, India's Digital Personal Data Protection Act (DPDPA), https://trustarc.com/resource/indias-digital-personal-data-protection-act-dpdpa/
[29] Copyright, AI, and the Future of Internet Search before the CJEU: Reflections on Like Company v Google, VerfBlog (July 17, 2025), https://verfassungsblog.de/copyright-ai-cjeu/, DOI: 10.59704/3887ddcde69e6944.
[30] Adam Buick, Copyright and AI training data—transparency to the rescue?, Journal of Intellectual Property Law & Practice, Volume 20, Issue 3, March 2025, Pages 182–192, https://doi.org/10.1093/jiplp/jpae102
[31] TrustArc, India's Digital Personal Data Protection Act (DPDPA), https://trustarc.com/resource/indias-digital-personal-data-protection-act-dpdpa/ (last visited Sept. 26, 2025)

automated copying for large-scale, commercial AI production is far removed from any legitimate analytic or research project.[32]

4.2 Implementation Challenges

The way TDM exceptions are implemented differs considerably across the EU, which creates legal uncertainty for cross-border AI production. The opt-out mechanism in Article 4 of the CDSM Directive requires "machine-readable" reservations, but technical standards for such mechanisms remain underdeveloped.[33] In a search engine indexing case, the recent Hungarian court decision suggests courts may interpret natural language opt-outs as sufficient. Still, interpretation conflicts with the directive's apparent requirement for technical implementation.[34]

The "lawful access" condition present in Articles 3 and 4 creates further complications.[35] While the condition ensures that copyright holders receive some compensation via access fees or subscription revenue to develop their AI, the condition may put small AI developers at a disadvantage simply because they may not be able to pay the very high licensing fees. This dynamic may benefit large technology companies.[36]

5. Recent Judicial Developments

5.1 United States Developments

The U.S. legal landscape has evolved; there have been a number of consequential developments in the U.S. legal field over this period of time, through a series of notable cases. Most significantly, the Authors Guild v. Google case that culminated in a Second Circuit decision in 2015 set critical precedents for massive digitization efforts.[37] The court's finding that Google's

[32] Rossana Ducato & Alain Strowel, Ensuring Text and Data Mining: Remaining Issues With the EU Copyright Exceptions and Possible Ways Out, CRIDES Working Paper No. 1/2021 (2021), https://verfassungsblog.de/wp-content/uploads/2021/04/Ducato_Strowel_CRIDES_WP_1_2021.pdf
[33] *The Exception for Text and Data Mining (TDM) in the Proposed Directive on Copyright in the Digital Single Market - Legal Aspects* (Jan. 2018), SSRN Electronic Journal, https://doi.org/10.2139/ssrn.3160586.
[34] Nicola Lucchi, Generative AI and Copyright: Training, Creation, Regulation, Policy Department for Citizens' Rights and Constitutional Affairs, European Parliament, PE 774.095 (2025), https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2025)774095
[35] Ibid
[36] Jiawei Zhang, Input out, output in: towards positive-sum solutions to AI-copyright tensions, Journal of Intellectual Property Law & Practice, Volume 20, Issue 9, September 2025, Pages 594–604, https://doi.org/10.1093/jiplp/jpaf037
[37] Petition for Writ of Certiorari, Authors Guild v. Google, Inc., No. 15-777 (U.S. Jan. 2016), https://www.scotusblog.com/wp-content/uploads/2016/01/Authors-Guild-v.-Google.pdf

book digitization program constituted fair use because it was "highly transformative" and provided "significant public benefits" has influenced subsequent AI training cases.[38]

A more recent case, Bartz v. Anthropic (2025)[39], is the first case in which a fair use defense for AI training on copyrighted books was successful. Judge Alsup focused more heavily on the transformative nature of using the books to train a language model and carefully pointed out that the AI system served different purposes than the books. The court was also careful to draw a line between lawful and unlawful sources, finding that training on pirated material weighed against a finding of fair use.[40]

The Authors Guild v. OpenAI [41] litigation is the latest and most thorough challenge to AI training practices. Featured in the case are allegations by famous authors, such as John Grisham and George R.R. Martin, that OpenAI's use of their works is beyond any fair use standard. The outcome will have concrete precedential value with respect to the commercial space for AI training.

5.2 Jurisprudence of the European Union

European courts have started to resolve the CDSM Directive's TDM provisions, with mixed outcomes for the purposes of AI developers. The most favorable outcome for AI developers comes from Germany's Kneschke v. LAION [42]case, where the Hamburg Regional Court found that research organizations could rely on Article 3 TDM exceptions even if the research ultimately leads to commercial applications.

However, the court's obiter dictum on Article 4 implies that commercial TDM might be under more scrutiny. The court was doubtful whether general opt-out notices in natural language would be effective machine-readable reservations under Article 4. This could strongly curb the generality of the TDM exception for commercial AI creation.

In the search engine indexing case, the Hungarian court decision further complicates the landscape by conflating web scraping with TDM. While the decision may expand the scope of

[38] Authors Guild v. Google, Inc., 954 F. Supp. 2d 282 (S.D.N.Y. 2013), aff'd, 804 F.3d 202 (2d Cir. 2015), cert. denied, 136 S. Ct. 1658 (2016).
[39] Bartz v. Anthropic PBC, No. 1:24-cv-00716 (N.D. Cal. 2025).
[40] Frank D'Angelo & Elena De Santis, Bartz v. Anthropic PBC, Loeb & Loeb LLP Client Alert (July 2025),
[41] Authors Guild v. OpenAI, No. 1:23-cv-08292 (S.D.N.Y. 2023).
[42] Kneschke v. LAION, Case No. 310 O 227/23, Hamburg Regional Court (Sept. 27, 2024)

TDM exceptions, it also creates uncertainty about the boundaries between different types of automated content processing.

5.3 Development in the Indian Legal Landscape

The ANI v. OpenAI case in the Delhi High Court represents India's first major confrontation with AI copyright issues. ANI seeks to establish direct copyright infringement due to OpenAI's unauthorized copying and misattribution of copyrightable work created by the AI. There are also jurisdictional complications in the case of OpenAI, insisting that there is no jurisdiction in India for work done overseas during training.[43]

The Delhi High Court has involved two amicus curiae experts - one in the field of intellectual property law and the other in copyright law - to illustrate the novel and complex legal questions involved in this case. The court's final decision may establish important precedent for the treatment of cross-border AI development activities under Indian law.[44]

Early proceedings indicate that Indian courts might be more restrictive than their U.S. equivalents. The court's emphasis on the commercial character of the activities of OpenAI and potential damage to the licensing market of ANI is an indication of skepticism in relation to extensive fair dealing defenses.

6. Database Rights and Sui Generis Protection

6.1 EU Database Directive Framework

The sui generis database right in the EU creates new challenges for training AI that do not exist elsewhere.[45] This right protects the investment of considerable labor to obtain, verify, or present a database to the public, regardless of whether the database is original enough to qualify for copyright protection.[46] This protection extends for 15 years from database completion, with potential for renewal through substantial changes.[47]

---

[43] Asian News International v. OpenAI Inc., W.P. (C) 123/2024, Delhi High Court (2024-2025)
[44] Ibiud
[45] Directive 96/9/EC of March 11, 1996, on the legal protection of databases, Official Journal L 77, 27.3.1996, p. 20–28.
[46] European Commission, Database protection in the EU, Your Europe Business portal (2021), https://europa.eu/youreurope/business/intellectual-property/database-protection/index_en.htm.
[47] CJEU, British Horseracing Board Ltd v William Hill Organization Ltd, Case C-203/02, ECLI:EU:C:2004:695 (July 8, 2004).

The Court of Justice of the European Union's interpretation in *British Horseracing Board* established that investment in data creation does not qualify for sui generis protection.[48] This distinction becomes crucial for AI-generated datasets, where the line between data creation and obtaining is often blurred. Recent academic analysis suggests that most machine-generated training datasets fall outside sui generis protection because they involve automated data creation rather than human curation.[49]

6.2 AI Training: Key Implications

The interaction of database rights and AI training gives rise to several intricate legal situations. AI developers may infringe Copyright (if the database qualifies for copyright protection due to originality) and/or sui generis rights (if they can show that such rights exist and a "significant" investment has been made in the database) if they take "a substantial part" of a database in order to train AI models. The "insubstantial parts" exception only permits minimal extraction, and even using "insubstantial parts" from multiple databases can constitute an infringement of one or both rights.[50]

The exceptions to the CDSM Directive's TDMs fail to reference database rights expressly, leaving doubt over their scope. Article 4 allows for copyrighted works to be reproduced for TDM use, but it is unclear whether this would include database extraction rights.[51] This legislative omission may greatly reduce the scope of TDM exceptions available for complete AI training processes.[52]

7. Licensing and Market Solutions

7.1 Voluntary Licensing Developments

The AI sector has started creating voluntary licensing frameworks to manage copyright and

[48] Gevers, S., The protection of modern databases under the sui generis database right, Gevers.eu Blog (Sept. 29, 2024), https://gevers.eu/blog/the-protection-of-modern-databases-under-the-sui-generis-database-right/.
[49] Transatlantic Lawyer, The Protection of Databases in the EU and under French Law (2021), https://www.transatlantic-lawyer.com/the-protection-of-databases-eu/.
[50] Lucchi, N., Generative AI and Copyright, European Journal of Risk Regulation 13(1), 54-98 (2022) (Analyzes the interplay between database rights and AI training).
[51] European Parliament, Navigating AI and IP: Dataset Use and Sui Generis Rights (2025), https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2025)774095 [Asian News International v. OpenAI].
[52] Directive 2019/790/EU (Digital Single Market Directive), arts. 3-4 (2019) (TDM exceptions, limited scope to copyright and unclear on database rights).

database right issues. Large publishing houses like The Financial Times, Associated Press, and news agencies have licensing contracts with AI firms.[53] Such contracts usually grant access to real-time as well as archived content in return for monetary compensation and attribution obligations.[54]

The development of collective licensing organizations for AI training in particular is an important market evolution.[55] Players such as the United States' Copyright Clearance Center and a range of European collective management societies are creating bulk licensing models of copyrighted work for AI uses. These programs seek to minimize transaction costs while protecting compensation for the rights holder.[56]

Nevertheless, voluntary licensing has several drawbacks. The global nature of AI innovation, combined with the enormous size of training datasets, makes total licensing unfeasible from an operational standpoint. In addition, differences in views on proper levels of compensation and the inherent difficulties in oversight of AI system outputs with regard to regulatory compliance persist as barriers.[57]

7.2 Statutory Licensing Proposals

Certain governments are considering statutory licensing systems to address voluntary licensing market failure. The United States Copyright Office Report in 2025 reviewed a number of licensing models, including extended collective licensing and compulsory licenses. However, the Copyright Office expressed skepticism about mandatory licensing models, noting concerns about government price-setting, as well as administrative burden.[58]

Extended collective licensing, by which representative organizations may license works on

[53] European Parliament, Generative AI and Copyright: Training, Use, and Licensing (2025), https://www.europarl.europa.eu/thinktank/en/document/IPOL_STU(2025)774095 [Asian News International v. OpenAI].
[54] International Federation of Reproduction Rights Organisations (IFRRO), Licensing Models for AI Training and Text/Data Mining (2023), https://ifrro.org/licensing-ai/.
[55] Authors Guild, Navigating Licensing Agreements in the Age of AI (2023), https://authorguild.org/ai-licensing-framework/.
[56] Marek, J., "The Economics of Licensing in AI Training," Journal of Intellectual Property Law & Practice, vol. 16, no. 7, pp. 549–560 (2021).
[57] Perel, M., and M. Roman, "Challenges of Collective Licensing for AI," European Intellectual Property Review, vol. 43, no. 5, pp. 287-298 (2021).
[58] U.S. Copyright Office, Copyright and Artificial Intelligence, Part 3: Generative AI Training (May 2025) (prepublication version), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-3-Generative-AI-Training-Report-Pre-Publication-Version.pdf

behalf of entire classes of rights holders, fared better. This mechanism, already employed in certain European countries for some other purposes, may offer a compromise between purely voluntary and wholly compulsory systems.[59]

India's technical committee also favors licensing-based options as one of its suggestions for Copyright reform. The panel's impending report, due in 2025, can suggest legislative changes to enable AI training, but with compensation to the rights holders.[60]

8. Cross-Border Enforcement Challenges

8.1 Jurisdictional Complexity

The international aspect of AI creation poses huge problems for enforcing copyrights. Training can take place in a country while the created AI systems are in use around the world, making it difficult to ascertain where infringement takes place. The ANI v. OpenAI case is one illustration of these issues, with OpenAI pleading that Indian courts would be without jurisdiction in cases of training taking place outside India.[61]

There is no worldwide harmonization on this matter, and the issue of forum shopping and harmonization still exists because certain jurisdictions are still traditional and have a territoriality-based conception of jurisdiction.

8.2 Technological Enforcement Measures

Technical measures implemented to prevent unauthorized AI training are evolving rapidly. Web scraping detection systems, robots.txt protocols, and other more complex blocking mechanisms give rights holders some limited level of protection. Yet, the arms race between scraping and anti-scraping technologies suggests that it is highly unlikely that any "technical" resolution will be able to independently resolve the legal disputes.[62]

---

[59] U.S. Copyright Office, Copyright and Artificial Intelligence, Part 1: Digital Replicas (July 2024), https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf

[60] U.S. Copyright Office Issues Report Addressing the Use of Copyrighted Material to Train Generative AI Systems," Squire Patton Boggs (July 2025), https://www.srz.com/en/news-and-insights/alerts/us-copyright-office-issues-report-addressing-use-of-copyrighted-material-to-train-generative-ai-systems.

[61] Andres Guadamuz, Artificial Intelligence and Copyright, 2024 WIPO Magazine, https://www.wipo.int/web/wipo-magazine/articles/artificial-intelligence-and-copyright-40141.

[62] Jin-Hee Lee, Dipunj Gupta, Brian Mitchell, Reid Tatoris & Henry Clausen, Control Content Use for AI Training with Cloudflare's Managed Robots.txt and Blocking, Cloudflare Blog (July 1, 2025),

The provision of "AI opt-out" mechanisms indicates a readiness to provide rights holders greater control over the use of their content. All the major AI actors accept the use of "opt-out" mechanisms, though in many jurisdictions, there is significant ambiguity as to whether they are legally enforceable. The CDSM's Article 4 EU requirement for rights holders to make available a machine-readable reservation may be a useful template for future use.[63]

9. Emerging Approaches in Policy Development

9.1 Regulatory Innovation

Some jurisdictions are either working toward or have established new models of regulation that see the goals of AI development and the protection of individual rights are balanced. The EU AI Act, which requires certain AI systems to be transparent, requires an AI system to disclose its copyrighted training data.[64] While this prospect may create a more definite path for rights holders to enforce rights in relation to the AI system, only time will tell if this is indeed a useful step in practice.[65]

Chinese AI regulation involves government control over the regulation of AI and over AI-produced content. There are also various provisions in Chinese law, including a uniquely Chinese model that is required when government approval is necessary for AI to train on specific types of content.[66] This model will not transfer to other jurisdictions, but it can offer some modeling of alternate approaches to AI development and regulation.[67]

9.2 International Harmonization Efforts

Global institutions are starting to tackle intellectual property concerns related to AI. While tangible results are few, the World Intellectual Property Organization has formed working

https://blog.cloudflare.com/control-content-use-for-ai-training/.

[63] Tiedrich, Lee, Perset & Fialho, Intellectual Property Issues in Artificial Intelligence Trained on Scraped Data, OECD Artificial Intelligence Papers No. 33, OECD Publishing, Paris (Feb. 2025), https://www.oecd.org/innovation/intellectual-property-issues-in-artificial-intelligence-trained-on-scraped-data.pdf.

[64] Office for Artificial Intelligence & Intellectual Property, European Parliament Research Service, AI Act: Legal and Regulatory Overview (2023).

[65] Kong, Ning & Weidl, Regina, Regulating AI in China: Government Controls and the Content Model, China Law Review (2024).

[66] Wang, Y., & Li, X., Chinese AI Regulation: Centralized Governance and Content Control, Asian Journal of Law and Technology, Vol. 32 (2023).

[67] OECD, Observations on AI Regulation in China, EU, and Other Jurisdictions, Policy Paper (2024).

groups for AI and IP. WIPO negotiations for trade agreements increasingly involve terms covering cross-border data flows and AI innovation.[68]

Global institutions are now beginning to tackle some of the intellectual property challenges posed by AI products. Although there aren't many initiatives with results, the World Intellectual Property Organization is now creating working groups on intellectual property and artificial intelligence. In a similar vein, cross-border data provisions and AI innovation are beginning to be included in many trade relations agreements.[69]

10. Critical Analysis and Recommendations

10.1 Assessment of Current Frameworks

Comparative analysis has shown that current legal frameworks have substantial weaknesses when it comes to addressing the issues of AI training data. The fair use doctrine in the US is somewhat flexible, but it carries a significant uncertainty that can easily suppress innovation or encourage abuse of rights. The more formalized regime inherent in the EU TDM exceptions is the most certain way to safely deploy AI training data in the short term, although it may prove to be too constraining to give legal assurances, the ability to keep pace with technological evolution, and the outdated institutional structures in India need a complete overhaul in order to be sufficiently contemporary to AI applications.[70]

No jurisdiction examined in this study appropriately addresses the relationship between copyright and data protection requirements. Interaction will become increasingly crucial as AI training incorporates personal data, intellectual property rights, and privacy protection. These issues are addressed as separate issues in current frameworks, which results in gaps and inconsistencies.[71]

---

[68] https://www.wipo.int/edocs/pubdocs/en/wipo-pub-econstat-wp-77-en-artificial-intelligence-and-intellectual-property-an-economic-perspective.pdf

[69] Alexander Cuntz, Carsten Fink & Hansueli Stamm, Artificial Intelligence and Intellectual Property: An Economic Perspective, WIPO Research Paper No. 77 (2024), available at https://www.wipo.int/edocs/pubdocs/en/wipo_pub_ecnstat_wp_77.pdf

[70] Markus Kattnig, Alessa Angerschmid & Thomas Reichel, Assessing trustworthy AI: Technical and legal perspectives of fairness in AI, 55 Computer L. & Sec. Rev. 106053 (2024), https://doi.org/10.1016/j.clsr.2024.106053.

[71] Pasetti, M., Santos, J.W., Corrêa, N.K. et al. Technical, legal, and ethical challenges of generative artificial intelligence: an analysis of the governance of training data and copyrights. Discov Artif Intell 5, 193 (2025). https://doi.org/10.1007/s44163-025-00379-6

10.2 Recommendations for Legal Reform

On this basis, a number of recommendations for enhancing legal frameworks are made:

- Global Standards for Harmonization: Global organizations must develop a model law related to rights surrounding AI training data, including Copyright and database protection. The standards should offer at least a minimal level of protection for rights holders in order to provide adequate exceptions for AI- generated content.

- Expanded Fair Use/Fair Dealing: Jurisdictions with restrictive fair dealing laws should consider expanding exceptions to cover transformative uses in training AI. Such expansion should, however, include provisions for rights holder remuneration and to avoid substitution in the market.[72]

- Compulsory licensing models: If voluntary licenses do not produce adequate solutions, consider compulsory licensing schemes. Extended collective licenses may offer the best combination of protection for rights holders and access.[73]

- Technical Standards for Opt-Out: Global development of technical standards for machine-readable reservation of rights would improve the efficiency of opt-outs while minimizing compliance costs.[74]

- Cross-Border Enforcement Mechanisms: International pacts must deal with issues of jurisdiction in AI training cases via mutual recognition of judgments as well as coordination of enforcement.[75]

11. Conclusion

The crossroads of machine learning data sets and intellectual property rights is one of the most complicated challenges faced by legal systems today. This comparative study finds that,

[72] Kretschmer, M., Margoni, T. & Oruç, P. Copyright Law and the Lifecycle of Machine Learning Models. IIC 55, 110–138 (2024). https://doi.org/10.1007/s40319-023-01419-3

[73] Monesh Mehndiratta, Concept of Compulsory Licensing: Indian Perspective, Ipleaders, https://blog.ipleaders.in/concept-compulsory-license-patents-act-1970/.

[74] Matthias Leistner, Lucie Antoine, TDM and AI Training in the European Union – From 'LAION' to Possible Ways Ahead?, GRUR International, 2025;, ikaf114, https://doi.org/10.1093/grurint/ikaf114

[75] Cross-Border Data Privacy and AI Governance: A Comparative Study Between the UK and the US, ResearchGate (year), https://www.researchgate.net/publication/395231846_Cross-Border_Data_Privacy_and_AI_Governance_A_Comparative_Study_Between_the_UK_and_the_US.

although each of the jurisdictions studied has sought to solve these challenges, no one has yet formulated a completely adequate system. The United States' flexible stance on fair use provides flexibility at the cost of certainty. The European Union's strict exceptions provide certainty, but may be too restrictive of high-technology advancement. The antiquated Indian template needs to be wholly reconsidered to retain any kind of relevance.[76]

Legal systems must strike a balance between two conflicting demands as a result of the growing need for AI development: on the one hand, the defense of legal intellectual property rights, and on the other, the growth of technology for the benefit of society. The emergence of voluntary licensing markets offers the potential of commercial solutions to help supplement optimal outcomes, but legal ambiguity stands in the way of such outcomes.[77]

Future evolution will involve coordinated global action to synchronize methods while being sensitive to policy agendas of other jurisdictions. The cases that are in the process of going through courts in each of the three jurisdictions will set key precedents, but finality could involve legislation and global cooperation.

Stakes lie much higher than technical legal issues. How these are solved will determine the evolution of AI technologies that have the potential to reshape human society in its very foundations. Achieving a balance between innovation and protection of rights is not just a legal issue but an important policy priority for the digital era.

While courts and legislatures struggle with these unprecedented issues, well-reasoned, evidence-based solutions are most important. The suggestions made in this examination set forth possible avenues forward, but all will need continued commitment from policymakers, technologists, and rights holders. The course of AI innovation and intellectual property protection lies within our collective capacity to build principles that advance the public interest while honoring valid private rights.

[76] Mohd Akhter Ali & M. Kamraju, Impact of Artificial Intelligence on Intellectual Property Rights: Challenges and Opportunities (Dec. 2023), available at https://www.researchgate.net/publication/376751087.
[77] Poonam Gulati & Megha Sharma, Intellectual Property at the Crossroads: AI and Copyright Concerns, in The Future of Law in a Globalized World: Navigating Innovation, Ethics and Sustainability, vol. 1, ISBN 978-81-986578-1-7 (May 2025).