International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 09 Issue: 07 | July - 2025 SJIF Rating: 8.586 ISSN: 2582-3930

Ethical and Legal Boundaries of Training AI on Protected Works: A Comprehensive Analysis of Intellectual Property Challenges in the Generative AI Era

Kinshuk Agarwal
Dept. of CSE
R. V. College of Engineering
Bengaluru, India
kinshukagarwal.cs22@rvce.edu.in

Kushagra Awasthi
Dept. of CSE
R. V. College of Engineering
Bengaluru, India
kawasthi.cs22@rvce.edu.in

LVS Aditya
Dept. of CSE
R. V. College of Engineering
Bengaluru, India
lvsaditya.cs22@rvce.edu.in

Mohammad Meezan
Dept. of CSE
R. V. College of Engineering
Bengaluru, India
mohammadmeezan.cs22@rvce.edu.in

Prof. Shruthi M N
Dept. of IEM
R. V. College of Engineering
Bengaluru, India
shruthimn@rvce.edu.in

Abstract—The exponential growth of generative artificial in-telligence has created unprecedented legal and ethical challenges regarding the use of copyrighted material in training datasets. This paper examines the complex intersection of intellectual property law and AI development, analyzing current litigation trends, platform policies, and regulatory responses. Through comprehensive analysis of recent lawsuits, including high-profile cases against OpenAI, Meta, and Anthropic, we investigate the boundaries between fair use and copyright infringement in AI training. Our research reveals that while AI companies claim fair use protection, courts are increasingly skeptical of blanket applications of this doctrine. The study examines platform responses such as Reddit and Twitter's API restrictions, the entertainment industry's collective bargaining efforts through SAG-AFTRA, and emerging legislative frameworks. We pro-pose a balanced approach that protects creators' rights while fostering innovation, including mandatory consent mechanisms, transparent attribution systems, and differentiated treatment for commercial versus non-commercial AI applications.

Index Terms—artificial intelligence, copyright law, fair use, intellectual property, generative AI, training datasets, digital rights

I. Introduction

The emergence of large language models (LLMs) and gener-ative artificial intelligence has fundamentally transformed the landscape of digital content creation and consumption. From OpenAI's GPT series to image generation models like Stable Diffusion and Midjourney, these systems have demonstrated remarkable capabilities in producing human-like text, images, and multimedia content [1]. However, this technological ad-vancement has precipitated a complex web of legal and ethical challenges centered on the use of copyrighted material in training datasets.

The scale of data consumption by modern AI systems is unprecedented. GPT-3, trained on approximately 570GB of text data representing trillions of words, exemplifies the massive datasets required for contemporary AI development [2]. This data encompasses books, articles, social media posts, images, videos, and other creative works, much of which is protected by copyright. The fundamental question emerges: does training AI systems on copyrighted material without explicit permission constitute fair use, or does it represent a systematic violation of intellectual property rights?

Recent legal developments have intensified this debate. High-profile lawsuits filed by The New York Times against OpenAI and Microsoft, music industry giants against AI music generators Suno and Udio, and visual artists against image generation platforms have brought these issues to the forefront of legal discourse [3]. Simultaneously, platforms like Reddit and Twitter have implemented API restrictions specifically to prevent unauthorized data mining for AI training, signaling a broader industry shift toward protecting user-generated content [4].

The entertainment industry has mobilized through collective bargaining, with the Screen Actors Guild-American Federa-tion of Television and Radio Artists (SAG-AFTRA) securing historic protections against unauthorized use of performers' likenesses in AI systems [5]. These developments reflect growing recognition that the current legal framework may be inadequate to address the unique challenges posed by AI training practices.

This paper provides a comprehensive analysis of the ethical and legal boundaries surrounding AI training on protected works. We examine the evolution of copyright law in the

context of emerging technologies, analyze current litigation trends, and evaluate platform and industry responses. Through detailed case studies and legal analysis, we aim to contribute to the ongoing discourse on balancing innovation with creator rights in the generative AI era.

II. Background and Literature Review

A. Evolution of Large Language Models

The development of large language models has followed a trajectory of exponential growth in both capability and data requirements. Early models like BERT (Bidirectional Encoder Representations from Transformers) and GPT-1 established the foundation for transformer-based architectures, training on relatively modest datasets [6]. GPT-1 utilized the BooksCorpus dataset, containing over 7,000 unique books, representing a significant but manageable corpus of training data.

The progression to GPT-2 and subsequently GPT-3 marked a paradigm shift in data scale and sources. GPT-3's training dataset, derived from Common Crawl, WebText2, and other internet-scale corpora, included content from diverse sources including news articles, academic papers, reference materials, and social media platforms [1]. This expansion raised imme-diate questions about the copyright status of training data and the implications of using copyrighted material without explicit licensing agreements.

B. Legal Framework for Copyright and Fair Use

Copyright law provides authors and creators with exclusive rights to reproduce, distribute, display, and create derivative works from their original creations. The fair use doctrine, codified in Section 107 of the U.S. Copyright Act, permits limited use of copyrighted material without permission for purposes such as criticism, comment, news reporting, teaching, scholarship, or research [7].

The application of fair use involves a four-factor analysis: (1) the purpose and character of the use, (2) the nature of the copyrighted work, (3) the amount and substantiality of the portion used, and (4) the effect of the use upon the potential market for or value of the copyrighted work. Courts have increasingly emphasized the transformative nature of use as a critical factor in fair use determinations [8].

C. Platform Terms of Service and Data Ownership

Major platforms hosting user-generated content maintain complex terms of service that govern data usage rights. YouTube's terms grant the platform broad rights to user-uploaded content for operational purposes but do not explicitly authorize third-party AI training [9]. Similarly, Reddit's up-dated developer agreement explicitly prohibits using platform data for training machine learning models without express permission [10].

The question of data ownership becomes particularly com-plex when considering user comments, posts, and interactions. While users retain copyright in their original creative works, platforms often claim licensing rights that may or may not extend to AI training applications. This ambiguity has created significant legal uncertainty and contributed to the current wave of litigation.

III. Current Legal Landscape

A. Major Copyright Litigation Cases

The legal landscape surrounding AI training has been shaped by several landmark cases that have established im-portant precedents and highlighted key areas of contention.

1) The New York Times v. OpenAI and Microsoft: Filed in December 2023, The New York Times lawsuit against OpenAI and Microsoft represents the first major media outlet to challenge AI companies' use of copyrighted content [3]. The complaint alleges that millions of Times articles were used to train GPT models without permission, creating systems that can reproduce Times content verbatim and compete directly with the newspaper's journalism.

The case has survived initial motions to dismiss, with a federal judge rejecting OpenAI's attempts to have the lawsuit thrown out. This represents a significant development, as it suggests courts may be willing to scrutinize AI companies' fair use claims more closely than previously anticipated.

2) Music Industry vs. AI Music Generators: In June 2024, major record labels including Sony Music, Universal Music Group, and Warner Music Group filed coordinated lawsuits against AI music generation companies Suno and Udio [11]. The complaints allege that these companies trained their mod-els on copyrighted music "at an almost unimaginable scale," enabling the generation of songs that closely mimic existing copyrighted works.

These cases are particularly significant because they focus on the creative process itself, arguing that AI-generated mu-sic directly competes with and potentially displaces human-created content in the marketplace.

3) Visual Artists' Collective Action: The visual arts com-munity has been at the forefront of copyright challenges to AI training. The case of Andersen v. Stability AI, filed in January 2023, involves a group of visual artists challenging the training of image generation models on copyrighted artwork without permission [12]. The case has survived multiple motions to dismiss and is proceeding toward discovery, potentially providing crucial insights into AI training practices.

4) Anthropic and Music Publishers: Concord Music Group and other major music publishers filed suit against Anthropic in October 2023, alleging that the company's Claude AI was trained on copyrighted song lyrics without permission [13]. The case highlights the particular vulnerability of lyrical con-tent, which is both highly copyrightable and easily identifiable when reproduced.

B. Recent Judicial Decisions

Recent court decisions have provided mixed signals regard-ing the application of fair use to AI training. In February 2025, a federal judge in the Thomson Reuters v. Ross Intelligence case rejected the defendant's fair use defense, marking the first major victory for copyright holders in an AI training context [14]. The decision emphasized that wholesale copying

of copyrighted databases for commercial AI development does not automatically qualify for fair use protection.

Conversely, other courts have been more receptive to AI companies' arguments. In the Anthropic case involving author claims, a federal judge ruled that using legally obtained books to train language models could constitute fair use, provided the use is transformative [15]. However, the judge also allowed the case to proceed to trial, indicating that factual questions remain regarding the specific nature and extent of the copying involved.

IV. Platform and Industry Responses

A. Social Media Platform Restrictions

Social media platforms have implemented increasingly re-strictive policies regarding data access and AI training, rec-ognizing both the commercial value of their content and the legal risks associated with unauthorized use.

1) Reddit's API Policy Changes: In April 2023, Reddit announced significant changes to its API pricing and usage policies, explicitly targeting AI training applications [4]. The new terms prohibit using Reddit data for training machine learning models without express permission and implement usage-based pricing that makes large-scale data extraction economically prohibitive for most AI companies.

Reddit's policy changes were motivated by recognition that user-generated content on the platform had significant commercial value for AI training. The company subsequently signed a licensing agreement with Google worth approx-imately $60 million annually, demonstrating the potential revenue streams available from controlled data licensing [16].

2) Twitter/X's Evolving Policies: Twitter, now known as X, has implemented a series of policy changes affecting AI training access. Initially, the platform introduced rate limits and API pricing changes that effectively restricted large-scale data extraction. More recently, X updated its privacy policy to allow sharing user data with third-party AI companies, while simultaneously prohibiting other companies from train-ing models on X content without authorization [17].

These seemingly contradictory policies reflect the platform's attempt to control and monetize its data while preventing unauthorized use by competitors. The changes have sparked significant user backlash and raised questions about user consent and data ownership.

B. Entertainment Industry Collective Bargaining

The entertainment industry has responded to AI challenges through collective bargaining and legislative advocacy, recog-nizing that individual creators lack the resources to challenge major tech companies independently.

1) SAG-AFTRA Strike and Agreement: The 2023 SAG-AFTRA strike marked a watershed moment in entertainment industry response to AI challenges. The 118-day strike, which ended in November 2023, resulted in historic protections for performers against unauthorized use of their likenesses in AI systems [5].

The final agreement established several key principles: (1) informed consent requirements for creating digital replicas of performers, (2) ongoing compensation for use of digital likenesses, and (3) restrictions on using existing footage to train AI systems without additional compensation. However, critics argue that the agreement's language contains loopholes that may allow studios to circumvent protections through careful contract drafting [18].

2) Writers Guild of America Protections: The Writers Guild of America secured complementary protections in their 2023 contract negotiations, establishing that AI-generated content cannot be considered "literary material" under guild agreements and that writers cannot be required to use AI tools in their work [19]. These protections reflect the guild's recog-nition that AI poses both a creative threat and an economic challenge to professional writers.

V. Case Studies

A. The Studio Ghibli Phenomenon

In March 2025, OpenAI's release of an enhanced image generation feature in ChatGPT led to a viral trend of users creating images in the distinctive style of Studio Ghibli animations [20]. The phenomenon raised complex questions about style copyright and the boundaries of permissible AI-generated content.

Studio Ghibli's co-founder Hayao Miyazaki had previously expressed strong opposition to AI in animation, stating in 2016 that he was "utterly disgusted" by AI-generated content
[21]. Despite this public stance, OpenAI's system clearly demonstrated the ability to replicate the studio's distinctive visual style, suggesting that Ghibli content had been included in training datasets without authorization.

Legal experts have suggested that Studio Ghibli could potentially pursue claims under the Lanham Act for trademark infringement and unfair competition, arguing that OpenAI's use of the "Ghibli style" creates consumer confusion about endorsement or licensing [?]. The case illustrates the particular challenges faced by creators whose distinctive styles become recognizable "brands" that AI systems can replicate.

B. DeepSeek and Alleged Data Distillation

The emergence of Chinese AI company DeepSeek in Jan-uary 2025 highlighted another dimension of AI training con-troversies: the practice of "distillation," where one AI model is trained using outputs from another [?]. OpenAI alleged that DeepSeek inappropriately used ChatGPT outputs to train its competing models, a claim supported by evidence that DeepSeek's system sometimes identified itself as ChatGPT.

The DeepSeek controversy raises several important legal questions. While model distillation is a common technique in AI development, using a competitor's outputs without per-mission may violate terms of service and potentially constitute unfair competition. The case also highlights the challenges of detecting and proving such practices, as distillation leaves few obvious traces in the final model [?].

C. Reddit v. Anthropic

In June 2025, Reddit filed a lawsuit against Anthropic alleging that the AI company had scraped Reddit content over 100,000 times without permission to train its Claude models [?]. The case represents a direct confrontation between a content platform and an AI company over data usage rights.

Reddit's lawsuit is particularly significant because it in-volves a platform that had explicitly prohibited AI training in its terms of service and had implemented technical measures to prevent large-scale data extraction. The case may establish important precedents regarding platforms' ability to control how their content is used for AI training.

VI. Technical and Economic Analysis

A. The Economics of AI Training Data

The economic dynamics of AI training data have shifted dramatically as the technology has matured. Early AI devel-opment benefited from the assumption that publicly available internet content could be freely used for research and de-velopment purposes. However, as AI systems have become commercial products generating billions in revenue, content creators and platforms have begun to assert economic claims to their data.

The mathematical relationship between data quality and model performance can be expressed through scaling laws. For language models, the relationship between performance and dataset size follows a power law:

$$L(D) = a \cdot D^{-\alpha} + L_{\infty} \quad (1)$$

where $L(D)$ represents model loss as a function of dataset size D, a and α are empirically determined constants, and L_{∞} represents the irreducible loss [?]. This relationship demonstrates why AI companies have strong incentives to access large, high-quality datasets, even when such access may involve legal risks.

B. Content Attribution and Provenance Tracking

Technical solutions for tracking content provenance in AI training datasets remain limited but are rapidly evolving. Blockchain-based systems for content authentication and wa-termarking technologies for AI-generated content represent potential approaches to addressing attribution challenges [?].

The implementation of content identification systems faces significant technical challenges. For text content, the relation-ship between input training data and model outputs is highly non-linear and difficult to trace. Research into "data influence" methods attempts to quantify how specific training examples affect model behavior:

$$I(x_i, \theta) = \nabla_{\theta} L(x_i, \theta) \cdot \nabla_{\theta} L(x_{test}, \theta) \quad (2)$$

where $I(x_i, \theta)$ represents the influence of training example x_i on the model's behavior for a test input, and L represents the loss function [?].

VII. Global Regulatory Responses

A. European Union Approach

The European Union has taken a proactive approach to AI regulation through the AI Act, which includes specific provi-sions addressing training data and copyright issues [?]. The legislation requires AI developers to implement "appropriate measures" to ensure that training data usage complies with copyright law, including maintaining detailed records of data sources and usage rights.

Under the EU framework, AI companies must demonstrate that they have either obtained proper licensing for copyrighted content or can justify their usage under applicable exceptions to copyright law. This represents a more restrictive approach than current U.S. practice and may influence global standards for AI development.

B. United States Legislative Developments

The U.S. Congress has introduced several bills addressing AI and copyright issues, including the NO FAKES Act, which would provide federal protection against unauthorized AI-generated replicas of individuals' voices and likenesses [?]. However, comprehensive federal legislation addressing AI training data remains under development.

State-level initiatives have been more aggressive. California has enacted legislation requiring specific consent for digital replica usage in employment contexts, while Tennessee has implemented the ELVIS Act protecting individual voice and likeness against unauthorized AI use [?].

VIII. Proposed Framework for Ethical AI Development

A. Consent and Attribution Mechanisms

Based on our analysis of current legal and industry de-velopments, we propose a multi-tiered framework for ethical AI training that balances innovation with creator rights. The framework incorporates the following key principles:

Informed Consent Requirements: AI developers should be required to obtain explicit consent for using copyrighted material in training datasets, with consent being specific to the intended use case and duration. This approach aligns with emerging industry practices and provides creators with meaningful control over their work.

Attribution and Revenue Sharing: When copyrighted content contributes to commercially successful AI systems, creators should receive appropriate attribution and, where applicable, revenue sharing. The technical implementation of such systems requires continued research into data influence methods and provenance tracking.

Differential Treatment by Use Case: The framework should distinguish between research, educational, and com-mercial applications of AI, with more permissive fair use applications for non-commercial research and stricter require-ments for commercial deployment.

B. Technical Implementation Standards

The proposed framework requires several technical capabilities that are currently under development:

1) Content Identification Systems: Robust methods for identifying copyrighted content in training datasets and tracking its contribution to model outputs.
2) Selective Training Exclusion: Technical capabilities to exclude specific content from training while maintaining model performance.
3) Post-Training Content Removal: Methods for remov-ing the influence of specific training examples from trained models when consent is withdrawn.

C. Economic Incentive Alignment

The framework must address the economic realities facing both AI developers and content creators. Revenue sharing mechanisms should be designed to:

- Provide fair compensation to creators based on their content's contribution to model capabilities
- Maintain incentives for AI innovation and development
- Establish sustainable business models for content licens-ing
- Support continued creation of high-quality training data

IX. Future Implications and Recommendations

A. Short-term Recommendations

Based on our analysis, we recommend the following immediate actions:

For AI Developers:

- Implement comprehensive content auditing systems to identify copyrighted material in training datasets
- Establish proactive licensing programs with content creators and platforms
- Develop technical capabilities for content attribution and influence tracking
- Engage in good-faith negotiations with creator communities and industry representatives

For Policymakers:

- Clarify fair use application to AI training through targeted legislation or regulatory guidance
- Support development of technical standards for content provenance and attribution
- Foster multi-stakeholder dialogue between AI developers, creators, and legal experts
- Consider international coordination on AI copyright standards

For Content Creators and Platforms:

- Develop clear policies regarding AI training permissions and restrictions
- Explore licensing opportunities that provide fair compensation while supporting innovation
- Implement technical measures to control unauthorized data access
- Engage in collective bargaining where appropriate to increase negotiating power

B. Long-term Considerations

The evolution of AI technology will continue to challenge existing legal and ethical frameworks. Several trends require ongoing attention:

Multimodal AI Systems: As AI systems increasingly integrate text, image, audio, and video capabilities, the complexity of copyright issues will expand. Future frameworks must account for the interactions between different types of copyrighted content.

Synthetic Data Generation: The development of AI sys-tems capable of generating high-quality synthetic training data may reduce reliance on copyrighted content but will raise new questions about data provenance and quality.

Federated Learning Approaches: Decentralized training methods that allow AI development without centralizing copyrighted content may provide technical solutions to some current challenges while introducing new governance complexities.

X. Conclusion

The intersection of artificial intelligence development and intellectual property law represents one of the most significant legal and technological challenges of our time. Our analysis reveals a rapidly evolving landscape where traditional copy-right concepts are being stress-tested by the unprecedented scale and nature of AI training requirements.

Current litigation trends suggest that courts are increasingly skeptical of blanket fair use claims by AI companies, particu-larly for large-scale commercial applications. The success of platform-imposed restrictions and industry collective bargain-ing demonstrates that technical and economic solutions can complement legal frameworks in protecting creator rights.

The proposed multi-tiered framework for ethical AI development recognizes that different use cases warrant different levels of protection and permission requirements. By incor-porating informed consent mechanisms, attribution systems, and revenue sharing where appropriate, this framework aims to balance the societal benefits of AI advancement with the legitimate rights of content creators.

However, significant challenges remain in implementing such frameworks. Technical limitations in content tracking and attribution, the global nature of AI development, and the rapid pace of technological change all complicate efforts to establish stable, effective governance systems.

The path forward requires continued collaboration between technologists, legal experts, policymakers, and creator communities. While the current period of uncertainty and litigation may be disruptive, it also presents an opportunity to establish precedents and frameworks that will guide responsible AI development for years to come.

As we move forward, the goal should not be to inhibit AI innovation but to ensure that such innovation occurs within a framework that respects the rights and contributions of the hu-man creators whose work makes these technological advances possible. The choices made in the coming years will determine whether AI development becomes a zero-sum competition

between technology and creativity or a collaborative endeavor that benefits both innovators and creators.

The stakes extend beyond immediate legal and economic concerns. The resolution of these issues will influence the future relationship between human creativity and artificial intelligence, the economic sustainability of creative industries, and the broader social acceptance of AI technologies. Getting this balance right is essential for realizing the full potential of artificial intelligence while preserving the human creativity that ultimately makes such systems valuable.

REFERENCES

[1] T. Brown et al., "Language models are few-shot learners," Advances in neural information processing systems, vol. 33, pp. 1877-1901, 2020.
[2] A. Radford et al., "Language models are unsupervised multitask learners," OpenAI blog, vol. 1, no. 8, p. 9, 2019.
[3] The New York Times Company v. Microsoft Corporation and OpenAI, Inc., Case No. 1:23-cv-11195, S.D.N.Y., Dec. 27, 2023.
[4] Reddit Inc., "API Update: Data API Access and New Features," Reddit Blog, April 18, 2023.
[5] Screen Actors Guild-American Federation of Television and Radio Artists, "Summary of the 2023 TV/Theatrical/Streaming Contracts," SAG-AFTRA, November 2023.
[6] J. Devlin et al., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," arXiv preprint arXiv:1810.04805, 2018.
[7] Copyright Act of 1976, 17 U.S.C. § 107 (1976).
[8] Campbell v. Acuff-Rose Music, Inc., 510 U.S. 569 (1994).
[9] YouTube, "Terms of Service," YouTube, updated June 1, 2023.
[10] Reddit Inc., "Reddit Developer Terms," Reddit, updated April 2023.
[11] Sony Music Entertainment v. Suno, Inc., Case No. 1:24-cv-04842, S.D.N.Y., June 24, 2024.
[12] Andersen v. Stability AI Ltd., Case No. 3:23-cv-00201, N.D. Cal., Jan. 13, 2023.
[13] Concord Music Group, Inc. v. Anthropic PBC, Case No. 3:23-cv-01092, M.D. Tenn., Oct. 18, 2023.
[14] Thomson Reuters Enterprise Centre GmbH v. ROSS Intelligence Inc., Case No. 1:20-cv-00613, D. Del., Feb. 11, 2025.
[15] Bartz v. Anthropic PBC, Case No. 3:24-cv-01034, N.D. Cal., Jan. 27, 2025.
[16] "Reddit signs content licensing deal with Google for AI training," Reuters, Feb. 22, 2024.
[17] X Corp., "Privacy Policy Update," X Platform, Nov. 15, 2024.
[18] K. Bond, "SAG-AFTRA's AI contract falls short on protections," Prism Reports, Dec. 5, 2023.
[19] Writers Guild of America, "Summary of the 2023 MBA," WGA, Sept. 27, 2023.
[20] "OpenAI's viral Studio Ghibli moment highlights AI copyright concerns," TechCrunch, Mar. 27, 2025.
[21] "Hayao Miyazaki reacts to AI animation," Great Big Story, Documentary footage, 2016.