University of Tehran Press

Comparative Law Review

Online ISSN: 2423-3404

Volume: 17, **Issue:** 1
Spring & Summer
2026

Homepage: https://jcl.ut.ac.ir

Copyright Challenges in Training Data for Generative AI Systems: A Study in US, EU, and Iranian Law

Pouyan Ahmadi[1] | **Ebrahim Rahbari**[2]✉

1. PhD Student in International Trade and Investment Law, Faculty of Law, Shahid Beheshti University, Tehran, Iran. Email: pouyan.ahm@gmail.com
2. Corresponding Author; Assistant Professor of International Trade Law and Intellectual Property and Cyberspace Law Department, Faculty of Law, Shahid Beheshti University, Tehran, Iran. Email: e_rahbari@sbu.ac.ir

Article Info

Article Type:
Research Article

Received:
2025/09/14
Received in revised form:
2025/12/16
Accepted:
2026/01/11
Published online:
2026/08/23

Keywords:
Copyright, Fair Use Exception, Generative Artificial Intelligence, Intellectual Property Law, Law of Emerging Technologies.

Abstract

1. Introduction

The rapid development of generative artificial intelligence has transformed contemporary technological and creative landscapes. Large Language Models (LLMs), including systems such as OpenAI's ChatGPT, are capable of generating human-like text, translations, summaries, computer code, and other forms of content through the analysis of enormous quantities of digital data. These systems rely on extensive datasets consisting of books, articles, websites, images, and other materials to identify linguistic and statistical patterns. As a result, the development of generative AI systems increasingly depends on the large-scale use of copyrighted works during the training process.

The growing reliance on copyrighted materials has generated serious legal challenges in the field of copyright law and intellectual property protection. The principal issue concerns whether the extraction, reproduction, storage, and processing of copyrighted works for the purpose of training AI systems constitute copyright infringement. This issue has become particularly significant because AI developers frequently collect data from online sources without obtaining explicit authorization from authors or copyright holders. Consequently, disputes have emerged regarding the legality of using protected works as training data.

Recent litigation in the United States illustrates the increasing importance of this issue. Cases such as *Tremblay v. OpenAI*, *Silverman v. OpenAI*, *The New York Times v. Microsoft and OpenAI*, and *Bartz v. Anthropic* have focused on allegations that copyrighted books, journalistic materials, and other protected works were used in AI training datasets without permission. These cases raise fundamental questions concerning the applicability of copyright exceptions and limitations to generative AI technologies.

At the international level, instruments such as the World Trade Organization TRIPS Agreement and the Berne Convention establish

Article Info

Abstract

general principles governing copyright limitations through the Three-Step Test. Nevertheless, domestic legal systems have adopted different approaches to the regulation of AI training data. The United States primarily relies on the judicial doctrine of fair use, while the European Union has adopted statutory exceptions for text and data mining (TDM) under the 2019 Directive on Copyright in the Digital Single Market. In contrast, Iranian copyright law does not contain explicit provisions governing AI training or data mining activities. Accordingly, this article examines the copyright challenges associated with generative AI training data through a comparative analysis of the legal systems of the United States, the European Union, and Iran.

2. Methodology

This study employs an analytical-descriptive methodology combined with a comparative legal approach. The descriptive aspect of the research explains the technical operation of generative AI systems and large language models in order to clarify the legal implications of AI training processes. The article examines how AI systems collect, reproduce, tokenize, and process digital content during machine learning operations.

The analytical and comparative dimensions of the study focus on the examination of copyright rules, judicial decisions, policy documents, and scholarly analyses in three legal systems:

1. **United States:** The study analyzes Section 107 of the U.S. Copyright Act and evaluates judicial interpretations of the fair use doctrine in recent AI-related litigation, including *Tremblay v. OpenAI*, *Silverman v. OpenAI*, *The New York Times v. Microsoft and OpenAI*, and *Bartz v. Anthropic*. The article examines how courts apply the four fair use factors to AI training activities.
2. **European Union:** The article analyzes Article 5(1) of the 2001 InfoSoc Directive concerning temporary acts of reproduction, as well as Articles 3 and 4 of the 2019 Directive on Copyright in the Digital Single Market relating to text and data mining for scientific and commercial purposes.
3. **Iran:** The study evaluates the Law for the Protection of Authors, Composers and Artists (1969), the Law on Translation and Reproduction of Books and Publications (1973), and recent Iranian AI policy documents in order to assess the legal status of AI training under Iranian copyright law.

Through this comparative framework, the study identifies the principal legal approaches, limitations, and regulatory gaps concerning copyrighted training data in generative AI systems.

3. Findings

A. The United States and the Doctrine of Fair Use

The findings demonstrate that the United States primarily addresses AI training through the doctrine of fair use under 17 U.S.C. §107. Courts evaluate fair use claims according to four factors: the purpose and

Article Info

Abstract

character of the use, the nature of the copyrighted work, the amount used, and the effect of the use on the market for the original work.

AI developers generally argue that training activities are transformative because copyrighted works are analyzed to identify patterns and relationships rather than reproduced for conventional expressive purposes. Recent judicial developments, however, indicate that the application of fair use to AI training remains uncertain and highly fact-dependent.

One important finding concerns the distinction between "input" and "output" infringement. In cases such as *Tremblay* and *Silverman*, some claims relating to AI-generated outputs were dismissed because plaintiffs failed to establish substantial similarity between the outputs and their original works. Nevertheless, allegations concerning the unauthorized copying of copyrighted materials during the training phase survived judicial scrutiny.

The study also finds that courts increasingly consider the source and method of data acquisition. In *Bartz v. Anthropic*, the court distinguished between legally obtained materials and unauthorized copies acquired from shadow libraries. The decision suggested that storing pirated copies of copyrighted works for AI training purposes may not qualify as fair use, even where the training activity itself possesses transformative characteristics. Accordingly, recent American jurisprudence demonstrates growing judicial attention to the legality of dataset acquisition in addition to the nature of AI training itself.

B. The European Union and Text and Data Mining Exceptions

Unlike the flexible American approach, the European Union has adopted a rule-based statutory framework through the 2019 Directive on Copyright in the Digital Single Market.

Article 3 of the Directive establishes an exception permitting research organizations and cultural heritage institutions to conduct text and data mining activities on lawfully accessible works for scientific research purposes. The exception allows the reproduction and extraction of copyrighted works without requiring prior authorization, provided that the data are securely stored and used exclusively for research-related purposes.

Article 4 extends text and data mining exceptions to commercial uses but grants copyright holders the right to reserve their rights through contractual terms, metadata, or machine-readable mechanisms. Consequently, commercial AI developers must verify whether rights holders have opted out of text and data mining permissions.

The findings indicate that although the European framework provides greater legal certainty than the American fair use doctrine, practical implementation challenges remain significant. In particular, the identification and enforcement of rights reservations present substantial difficulties because large-scale web scraping operations frequently involve massive quantities of online data originating from

Article Info

Abstract

numerous sources. The article therefore concludes that the effectiveness of the European opt-out mechanism remains uncertain in practice.

C. Iranian Copyright Regime and the Existential Legislative Gap

The study finds that Iranian copyright law does not contain explicit provisions addressing generative AI systems, text and data mining, or computational uses of copyrighted materials. Existing legislation, particularly the Law for the Protection of Authors, Composers and Artists (1969), primarily focuses on protecting the exclusive rights of authors and creators.

Although Iranian law recognizes limited exceptions for educational use, private use, and quotation, these exceptions are narrow in scope and do not address the large-scale reproduction and processing of copyrighted works required for AI training. Unlike the United States, Iranian law does not recognize a general doctrine comparable to fair use. Similarly, no statutory text and data mining exceptions comparable to those of the European Union exist within the Iranian legal system.

The article concludes that the use of copyrighted works in AI training under current Iranian law likely conflicts with the exclusive reproduction rights granted to copyright holders. The absence of clear legislative guidance has therefore created significant legal uncertainty regarding the development and training of AI systems in Iran. Recent Iranian policy documents relating to artificial intelligence have acknowledged the necessity of reforming intellectual property regulations and facilitating lawful access to data for AI development, but comprehensive legislative reform has not yet occurred.

4. Conclusion

The comparative analysis demonstrates that traditional copyright frameworks face substantial challenges in addressing the technological realities of generative artificial intelligence. The United States relies on the flexible doctrine of fair use, allowing courts to adapt copyright principles to new technological contexts, although this approach also creates uncertainty due to case-by-case judicial interpretation. The European Union, by contrast, has adopted detailed statutory text and data mining exceptions that provide greater legal certainty but also create implementation difficulties concerning rights reservations and commercial data mining activities.

Iranian copyright law currently lacks explicit legal mechanisms governing AI training and text and data mining activities. Existing statutory provisions were developed prior to the emergence of generative AI technologies and therefore do not adequately address contemporary technological developments. As a result, the use of copyrighted works for AI training purposes remains legally uncertain under Iranian law.

Article Info	**Abstract**
	The study concludes that balancing technological innovation with the protection of copyright holders requires clearer and more coherent legal frameworks. The article emphasizes the importance of establishing legal mechanisms capable of supporting scientific and technological development while simultaneously safeguarding the economic and moral rights of authors and creators. Potential approaches discussed in the article include the development of open-access datasets, licensing agreements between AI developers and rights holders, compensation mechanisms for creators, and clearer statutory regulation of text and data mining activities.
How To Cite	Ahmadi, Pouyan; Rahbari, Ebrahim (2026). Copyright Challenges in Training Data for Generative AI Systems: A Study in US, EU, and Iranian Law. *Comparative Law Review*, 17 (1), 25-50. DOI: https://doi.com/10.22059/jcl.2026.400962.634797
DOI	10.22059/jcl.2026.400962.634797
Publisher	