

Abstract

Designing a Limited Text and Data Mining Exception for Generative AI Training

Dae-Hee Lee*

Whether the use of copyrighted works for generative AI training constitutes fair use remains uncertain, discouraging both copyright enforcement and lawful data use. Statutory licensing and levy systems may impose remuneration even on potentially fair uses and require complex compensation mechanisms, while extended collective licensing can address only limited categories of works.

This paper proposes a limited text and data mining (TDM) exception for uses highly likely to qualify as fair use. Research-oriented TDM should be broadly permitted subject to lawful access, while general generative AI training may additionally require lawful sources and reasonable measures to reduce the reproduction of protected expression. Noncommercial purpose, rights reservation, non-enjoyment purpose, and unreasonable prejudice to copyright holders' interests should not operate as independent statutory requirements. The proposed exception would create a limited legal safe harbor rather than a blanket exemption.

Keywords

generative artificial intelligence, AI training data, text and data mining, fair use, copyright exception, extended collective licensing, opt-out

* Professor, School of Law, Korea University

참고문헌

1. 논문

- Brown, Tom B. et al., Language Models Are Few-Shot Learners, 33 Advances in Neural Information Processing Systems 1877 (2020).
- Calabresi, Guido & Melamed, A. Douglas, Property Rules, Liability Rules, and Inalienability: One View of the Cathedral, 85 Harv. L. Rev. 1089 (1972).
- Dou, Guangyao et al., Avoiding Copyright Infringement via Large Language Model Unlearning, in Findings of the Association for Computational Linguistics: NAACL 2025 5191 (2025).
- Fisher, William W. III, The Implications for Law of User Innovation, 94 Minn. L. Rev. 1417 (2010).
- Frosio, Giancarlo, Should We Ban Generative AI, Incentivise It or Make It a Medium for Inclusive Creativity?, in A Research Agenda for EU Copyright Law 61 (Enrico Bonadio & Caterina Sganga eds., Edward Elgar Publishing 2025).
- Geiger, Christophe/Iaia, Vincenzo, The Forgotten Creator: Towards a Statutory Remuneration Right for Machine Learning of Generative AI, 52 Comput. L. & Sec. Rev. 105925 (2024).
- Goldstein, Paul, Preempted State Doctrines, Involuntary Transfers and Compulsory Licenses: Testing the Limits of Copyright, 24 UCLA L. Rev. 1107 (1977).
- Keaton, Ashlye M./Goolsby, Jerry, In the Trenches of Copyright Law: Challenges and Remedies, 12 Tul. J. Tech. & Intell. Prop. 211 (2009).
- Lee, Edward, Copyright Dilution Under Constitutional Scrutiny, 25 Chi.-Kent J. Intell. Prop. 1 (2026).
- Lee, Katherine et al., Deduplicating Training Data Makes Language Models Better, in Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics 8424 (2022).
- Lemley, Mark A./Weiser, Philip J., Should Property or Liability Rules Govern Information?, 85 Tex. L. Rev. 783 (2007).
- Lewis, Patrick et al., Retrieval-Augmented Generation for Knowledge-Intensive