

AI AND COPYRIGHT: FROM A DOCTRINAL CRISIS TO A HYBRID MODEL OF COLLECTIVE LICENSING

Oleksii Shamov

Human Rights Educational Guild
18001, 40/28 Rizdvyana, Cherkasy, Ukraine
e-mail: yursprava@gmail.com
<https://orcid.org/0009-0009-5001-0526>

Abstract

The explosive development of generative artificial intelligence (hereinafter – AI) has triggered a fundamental crisis within intellectual property law. A central conflict has emerged between artificial intelligence developers, who require access to vast datasets for model training, and rights holders, who seek to protect their works and receive fair compensation for their use. The purpose of this article is to demonstrate the inadequacy of existing legal doctrines in addressing the use of copyrighted data for training generative AI and to substantiate the need for a transition to a hybrid model of compulsory collective licensing as the most rational alternative. The research is based on a comprehensive application of general scientific and special legal methods. The formal-legal method was used to analyze the norms of United States law and European Union (hereinafter – EU) law. The comparative-legal method allowed for the identification of common features and key differences between the American and European approaches and an assessment of their relative effectiveness. Systemic analysis was applied to view copyright as an integral system and to identify the fundamental challenges posed by artificial intelligence to its core elements (authorship, originality, economic incentives). The synthesis method enabled the generalization of the findings and the formulation of the author's hypothesis. The article proves that neither the flexible but unpredictable fair use doctrine nor the formalized but ineffective Text and Data Mining (hereinafter – TDM) exception can resolve the issue. As a solution, the architecture of a hybrid model of compulsory collective licensing is proposed. This research affirms that a continued reliance on legacy legal doctrines creates a landscape of perpetual instability. The case-by-case litigation engendered by the fair use doctrine proves to be an inefficient and unpredictable means of governance, while the EU's TDM opt-out mechanism has been shown to be functionally incapable of providing meaningful compensation to creators.

Keywords: artificial intelligence; collective licensing; copyright; fair use; text and data mining.

1. Introduction

We are witnessing a “Cambrian explosion” in the field of AI. Generative models such as OpenAI's GPT-4, Midjourney, and Stable Diffusion demonstrate a remarkable ability to create complex texts, photorealistic images, music, and software code, previously considered the exclusive domain of human intellect. This technological leap, which promises unprecedented opportunities for science, business, and art, has simultaneously precipitated a profound crisis in one of the most conservative areas of law-copyright (Eichner, 2023). The system, built over centuries to protect and incentivize human creativity, has proven unprepared for the emergence of a non-human “creator” capable of learning from the entirety of human knowledge and generating derivative content on an industrial scale.

At the heart of the conflict lies a fundamental dilemma. Training modern generative models requires enormous amounts of data (datasets) that inevitably contain billions of copyrighted texts and images scraped from the Internet. From the perspective of AI developers, this process is a necessary condition for innovation and falls under the concept of “non-expressive use”, as the model learns statistical patterns rather than copying specific works (Sag, 2024). On the other hand, writers, artists, photographers, and publishers see this as unauthorized mass copying that devalues their work and creates a direct competitor trained on their own intellectual property. This tension has already erupted into a series of high-profile lawsuits that have become markers of the era. Lawsuits from Getty Images, a group of artists led by Sarah Andersen, the Authors Guild, and especially the lawsuit by The New York Times against OpenAI and Microsoft, have moved theoretical discussions into the realm of multi-billion dollar legal battles (United States District Court for the Southern District of New York, 2024a, 2024b).

The issue of the interaction between AI and copyright is being actively researched by leading scholars worldwide. Recent publications analyze various aspects of this complex relationship. For instance, Abbott (2020) puts forward the radical idea of the need to grant AI certain rights, including copyright, to stimulate innovation. Samuelson (2023), one of the most authoritative researchers in the field of IP law, analyzes the applicability of the fair use doctrine, noting its flexibility but also its unpredictability. Researchers in the EU, as seen in a report for the European Parliament, are focused on analyzing the effectiveness of the Directive on Copyright in the Digital Single Market (hereinafter – CDSM), and are increasingly concluding that it is limited and that licensing solutions are needed (Lucchi, 2025). Gervais (2020) examines the problem through the prism of the fundamental blurring of the line between idea and expression, which is a cornerstone of copyright law.

Despite the significant number of publications, most of them focus either on analyzing specific aspects of the problem (authorship, fair use) or on criticizing existing approaches. A key part of the problem remains unresolved: the development of a comprehensive, practically feasible, and fair model that would not just state the crisis but would offer a systemic way out of it. This article is dedicated to this unresolved part, proposing a specific architecture for a hybrid model of collective licensing that synthesizes the best elements of global experience and adapts them to the unique challenges posed by generative AI. Thus, the objective of the research is not only to critique existing doctrines but also to substantiate and provide a detailed description of a viable alternative capable of creating legal certainty and ensuring the sustainable development of both the technological and creative spheres.

After outlining the rise of generative AI and the legal uncertainties it introduces, the scholarly conversation has fractured into several key debates. The scientific discourse in recent years surrounding AI and copyright has been characterized by high intensity and polarization of opinions. An analysis of publications in leading law journals indexed in the Scopus and Web of Science databases allows us to identify several key areas of research. The first area, represented by authors such as Sag (2024), focuses on attempting to adapt the fair use doctrine to the process of AI training. The argument is based on the idea that using data for training is “transformative” and “non-expressive”, as the model does not store or reproduce works but only learns statistical patterns.

However, as opponents note, this logic collapses when the generated content begins to compete directly with the original works on the market, which is a central issue in the NYT v. OpenAI case. The second area, more characteristic of European legal thought, analyzes the effectiveness of the TDM exception (De Pauw & Heremans, 2024; Lucchi, 2025). Researchers conclude that the right to opt-out, granted to rights holders by Article 4 of the CDSM Directive, is weak and technically difficult to implement, making the exception practically ineffective for protecting rights. The third, more philosophical direction, represented by the works of Abbott (2020) and others, questions the very

anthropocentrism of copyright law. The previously unresolved part of the general problem, to which this article is dedicated, is the absence in the scientific discourse of a detailed and substantiated model that would go beyond the dichotomy of “allow everything” (through broad interpretation of exceptions) or “prohibit everything” (through strict protection). Most studies either criticize the current state of affairs or propose piecemeal changes. This work, in contrast, proposes to consider the problem not as a matter of exceptions, but as a matter of mass commercial use that requires a systemic licensing solution, and describes the architecture of such a model in detail.

The relevance of this research lies in the fact that legal uncertainty is hampering the development of both sides. AI developers operate under the Damoclean sword of potential court injunctions and massive fines, which stifles investment. Creators, in turn, lack effective mechanisms to control the use of their works and receive fair compensation, undermining the economic foundations of the creative industries.

The purpose of this article is to prove that attempts to adapt existing doctrines, particularly the American fair use doctrine and the European exception for TDM, are futile. These tools, designed for entirely different realities, are incapable of providing an adequate balance of interests. Instead, the article argues for the necessity of a paradigm shift a transition to a hybrid model of compulsory collective licensing as the only rational way out of the impasse.

2. Materials and Methods

This study is a theoretical legal investigation based on the analysis of doctrines, legislation, and judicial practice. Given the nature of the objectives, a system of general scientific and special legal methods was used, which ensured the comprehensiveness and objectivity of the conclusions. The foundation of the methodology is the formal-legal method. With its help, a detailed analysis of the legal norms governing copyright was conducted, particularly the provisions of the U.S. Copyright Act, especially Section 107 on fair use, and the Directive (EU) 2019/790 on copyright and related rights in the Digital Single Market, specifically its Articles 3 and 4 concerning the exception for TDM. The analysis was aimed at identifying the content, structure, and limits of the application of these norms in the context of generative AI.

A key role in the research was played by the comparative-legal method. By comparing the American approach, based on a flexible judicial doctrine, and the European approach, based on codified legislative exceptions, their strengths and weaknesses were identified. This method allowed not only to state the differences but also to assess the potential effectiveness of each approach in achieving a balance of interests between AI developers and rights holders, which became the basis for the conclusion about the inadequacy of both systems.

Systemic analysis was applied to view copyright not as a set of disparate norms, but as an integral system that performs certain functions, primarily stimulating creativity by granting temporary monopoly rights. It was analyzed how the emergence of AI affects the key elements of this system: the concept of “author”, the criterion of “originality”, the “idea/expression” dichotomy, and the economic balance. This allowed us to go beyond a purely technical analysis of norms and understand the deep, fundamental nature of the crisis.

The final stage was the synthesis method, which consisted in generalizing the results obtained using the previous methods. Based on the identified shortcomings of existing legal models and taking into account the successful experience of collective rights management mechanisms in other areas (e.g., the music industry), the author's hypothesis – a hybrid model of compulsory collective licensing – was formulated and substantiated.

The materials (research base) for the work were: texts of legal acts of the USA and

the EU; materials of current court cases, in particular, statements of claim and decisions of first-instance courts in the cases *New York Times Company v. Microsoft Corporation*, et al., *Andersen v. Stability AI Ltd.*, *Authors Guild v. OpenAI, Inc.*; as well as scientific publications (articles, monographs, reports) by leading world researchers in the field of intellectual property law, published mainly in peer-reviewed journals indexed in international scientometric databases Scopus and Web of Science. An empirical part in the form of sociological surveys or experiments was not envisaged in this study.

3. Results and Discussion

3.1. The failure of fair use and the TDM exception

The main result of the research is the substantiation of the thesis that existing legal paradigms are incapable of adequately responding to the challenges of generative AI, and the only way out is to create a new system of regulation. This conclusion is based on a consistent analysis of fundamental problems and existing doctrines.

The copyright crisis caused by AI has two main dimensions: the problem of authorship for generated content and the problem of the legality of using protected data for training models. The first dimension, although important, is less acute. The position of the U.S. Copyright Office (2023) is that protection is granted only to works created by a human, and content fully generated by AI is not subject to protection. This provides some clarity, although it leaves open the question of works created in collaboration between a human and AI. However, the second problem – the legality of AI training – is much more complex and central. It is here that existing doctrines demonstrate their complete failure (Lemley & Casey, 2021).

The fair use doctrine, enshrined in Section 107 of the U.S. Copyright Act, is a flexible standard that allows courts in each specific case to determine whether the use of a work without the author's permission was fair. Courts analyze four factors: (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.

AI developers build their defense primarily on the first factor, arguing that training models is a “transformative use”. They refer to the decision in *Authors Guild v. Google, Inc.*, where the court recognized the creation of the Google Books digital library as fair use because the purpose (creating a search tool) was new and did not compete with the sale of books (Lemley & Casey, 2021). By analogy, they argue that AI does not store copies but only learns patterns, creating a new, transformative technology.

However, this argument begins to crumble when analyzing the fourth factor – the effect on the market. As the *New York Times* lawsuit convincingly demonstrates, generative models are capable of reproducing significant portions of original articles almost verbatim, providing users with answers to queries and thereby eliminating the need to visit the publication's website. This is no longer “non-expressive use”, but direct market substitution. When an AI assistant, trained on millions of recipes from culinary websites, generates its own recipe, it directly competes with these sites for attention and advertising revenue. The *Andersen v. Stability AI* case also highlights that AI trained on the works of specific artists can generate images “in the style” of that artist, which directly devalues their unique market proposition (United States District Court for the Northern District of California, 2024).

Moreover, the third factor (the amount of use) also works against AI developers. Unlike Google Books, where snippets were used, effective training of generative models requires full texts and images, meaning 100% of the work is copied, and billions of times over. Thus, reliance on fair use creates a situation of legal roulette. The outcome depends on the composition of the court, the specific facts, and the economic arguments that the parties can present. This leads to long, expensive court proceedings that cannot give a clear signal to the market. This doctrine is too unpredictable to serve as a reliable

basis for a multi-billion dollar industry.

The EU tried to approach the problem more systematically by introducing a special exception for TDM in the 2019 CDSM Directive. Article 3 provides a broad exception for TDM for the purpose of scientific research (Margoni & Kretschmer, 2022). However, for commercial purposes, which is relevant for most AI developers, Article 4 applies. It allows free TDM, but with one significant reservation: rights holders can expressly reserve their rights against such use (make an “opt-out”), for example, in a machine-readable format (e.g., in the robots.txt file).

At first glance, this looks like a balanced solution. However, in practice, it has proven ineffective for several reasons. First, for an AI developer who wants to train a model on data from the entire Internet, it becomes technically impossible to track the “opt-out” status for billions of individual web pages, images, and texts. Second, the mechanism is not unified, which creates fragmentation (Kop, 2021). Third, and most importantly, the model does not provide for any compensation for rights holders whose data were still used (if they did not opt-out). It puts them before a choice: either completely prohibit the use and lose any potential benefit, or allow it for free. As noted in the report for the European Parliament, this approach does not create a fair market for data, but only provokes a zero-sum game (Lucchi, 2025).

This analysis reveals that both approaches, despite their different philosophies, converge on a similar outcome of insufficiency. The American path leads to a costly and unpredictable landscape of litigation, while the European path results in a fragmented and economically unsatisfactory regulatory framework. Neither provides a viable foundation for the future. It is worth noting that the search for a balance of interests is not limited to the US and the EU. Other leading jurisdictions are also actively shaping their own approaches, which further underscores the global nature of the problem. For example, Japan holds one of the most liberal positions. Article 30-4 of its copyright law permits the use of copyrighted works for “information analysis” (which includes AI training) without the rights holder's permission, provided it does not “unreasonably prejudice” their interests. This flexible wording, however, creates its own legal uncertainty as to what exactly constitutes “unreasonable prejudice”.

The United Kingdom, which has its own TDM exception for non-commercial research, considered introducing a broader exception for commercial purposes but abandoned the idea after significant opposition from the creative industries. This demonstrates a deep conflict of interest that cannot be resolved by simply expanding exceptions. Meanwhile, Canada is still in the public consultation phase, analyzing various policy options. This mosaic of divergent approaches – from Japan's liberal stance to the UK's cautious one – demonstrates the absence of a single, universally accepted solution and strengthens the argument for developing a new, harmonized international model that could offer universal principles rather than a set of fragmented national rules.

3.2. The way forward: A hybrid model of compulsory collective licensing

The failure of existing legal doctrines creates not only a legal vacuum but also a huge economic risk, stifling the development of a market valued in the billions of dollars. According to a report from Bloomberg Intelligence, the generative AI market, valued at \$40 billion in 2022, is projected to grow to \$1.3 trillion by 2032. This exponential growth is based on the use of data, a significant portion of which is protected by copyright. However, unlike other industries where the mass use of content generates substantial licensing streams, the AI field currently remains a “wild west”.

For comparison, the global music industry's revenues from licensing (including public performance, synchronization, and other rights) amount to billions of dollars annually. Just two leading American collective management organizations, ASCAP and BMI, together collected over \$3 billion in licensing royalties for their authors and publishers in 2023 (Catsaros, 2023). This experience proves that mature and transparent

collective licensing systems can effectively monetize the mass use of content, providing a stable income for creators. Ignoring such an approach for the generative AI industry means losing potential billions in revenue for the creative economy and perpetuating a state of legal uncertainty, which is disadvantageous for all market participants in the long run.

It follows that licensing the mass commercial use of protected works for AI training is an adequate way to regulate these relations. Individual licensing (when the AI developer enters into agreements with millions of individual authors) is absolutely impossible due to the huge transaction costs (Pope, 2024). Therefore, the only realistic solution is a model of mandatory collective licensing. Since existing copyright exceptions are inadequate, the logical conclusion is that the mass commercial use of protected works for AI training should not be an exception. It should be regulated through licensing. However, individual licensing (where an AI developer enters into agreements with millions of individual authors) is absolutely impossible due to huge transaction costs. The only realistic solution is a model of compulsory collective licensing.

The idea of collective rights management is not new. It has been successfully operating for decades in other areas where individual licensing is impractical. A classic example is the music industry, where organizations such as ASCAP and BMI in the USA collect royalties from radio stations, television channels, and public establishments for the use of music and distribute them among authors and publishers. Another example is the Copyright Clearance Center, which provides licenses to universities and companies for photocopying scientific articles and books.

The proposed model for AI is based on the same logic: if a certain use is socially useful and economically important, but impossible on the basis of individual agreements, the state can establish a system of compulsory licensing that guarantees both access for users and compensation for creators (Hilty et al., 2020). The model should be hybrid, combining elements of the European tradition of collective management and American pragmatism. Its architecture could look as follows:

1. Subjects:

□ Payers: Developers of commercial generative AI models (e.g., OpenAI, Google, Stability AI).

□ Recipients: Rights holders (writers, artists, publishers, musicians, photographers).

□ Intermediary: One or more state-authorized collective management organizations (CMOs) operating on a non-profit basis.

2.Object of Licensing: The use of copyrighted works as part of a training dataset for a commercial generative AI model.

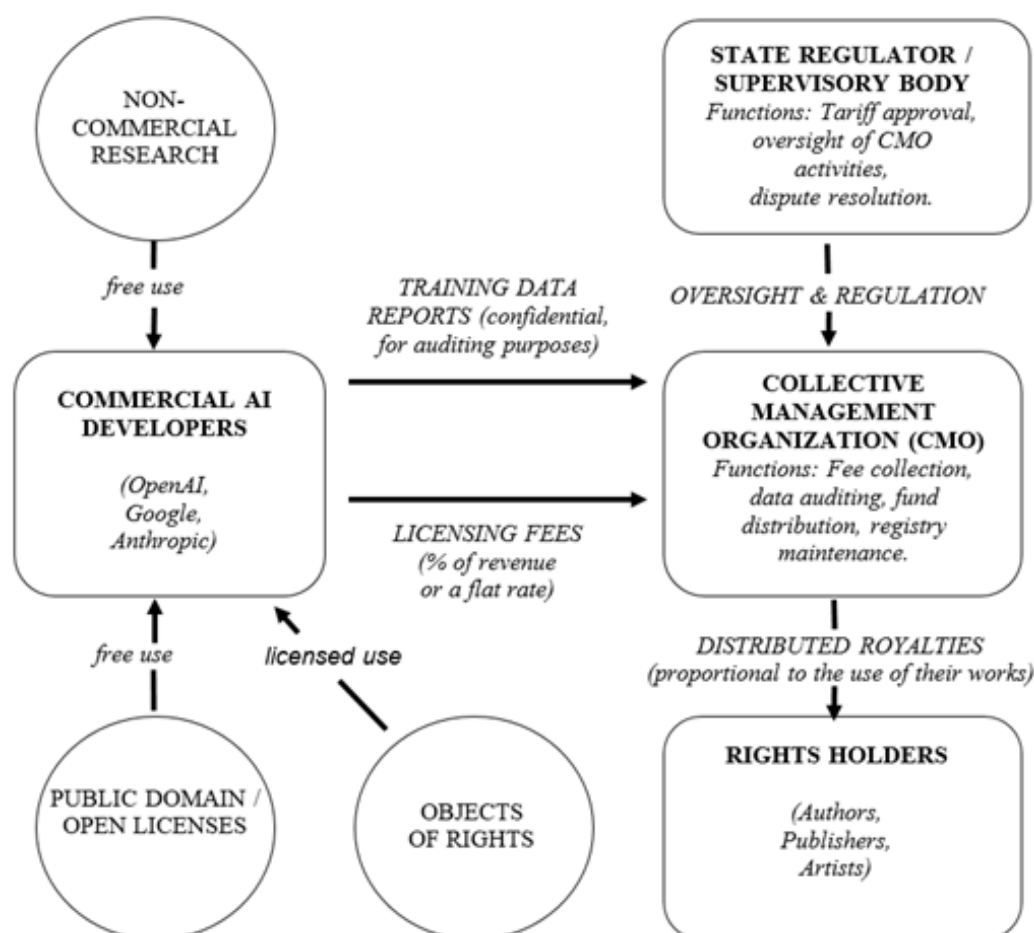
3.Mechanism of Operation:

□ Compulsory Payments: Legislation establishes an obligation for developers of commercial AI to pay license fees to the CMO. The payment can be structured as a percentage of the revenue received from the model, or as a fee that depends on the size and composition of the training dataset. Transparency Requirement: AI developers are obliged to provide the CMO with detailed but confidential reports on the composition of their training data. This is a key requirement, without which fair distribution is impossible. Technically, this can be implemented, and initiatives such as the Coalition for Content Provenance and Authenticity (C2PA) are already developing standards for tracking content provenance.

□ Distribution of Funds: The CMO, based on the reports received and with the help of its own analytical tools, distributes the collected funds among the rights holders in proportion to the share of their works in the training datasets.

□ Exceptions: The model does not apply to the use of data for non-commercial scientific research (analogous to Article 3 of the CDSM Directive) and to the use of works in the public domain or distributed under free licenses (e.g., Creative Commons).

Figure 1 illustrates the architecture of the Proposed Hybrid Model:



The proposed model has a number of key advantages:

- Legal Certainty: AI developers get a clear and legal way to access data, which eliminates the risk of lawsuits and promotes the stable development of technology.
- Fair Compensation: Creators receive remuneration for the use of their work, which supports the viability of the creative industries.
- Efficiency: The model eliminates the huge transaction costs associated with both individual licensing and litigation.

The implementation of such a model is associated with challenges. The main one is the need for international cooperation, as AI development and the Internet are global. However, this can be resolved through bilateral and multilateral agreements between national CMOs, as is already happening in the music sphere. The technical complexity of data auditing is significant but a solvable problem in the era of big data. Determining fair tariffs will require negotiations between all interested parties under the supervision of state regulators. Ultimately, the transition to such a model is not just one of the options, but a necessity. It is the only way to move beyond the “allow all” or “prohibit all” paradigm and create a balanced ecosystem in which AI innovation and human creativity can not conflict, but mutually complement and enrich each other.

4. Conclusions

The conducted research allows us to conclude that there is a deep systemic crisis in copyright law caused by the rapid development of generative artificial intelligence. Attempts to apply existing legal doctrines, developed for the analog era, to new technological realities have proven to be a failure.

It has been demonstrated that the American fair use doctrine, despite its flexibility,

is unable to provide the necessary legal certainty. Its application in the context of AI training leads to lengthy, expensive, and unpredictable court proceedings, where the outcome depends on the subjective assessment by the court of the balance of the four factors. This creates an unstable environment for both technology developers and rights holders.

Similarly, the European approach in the form of the exception for TDM, although an attempt at a codified solution, has proven ineffective in practice. The opt-out mechanism is technically difficult for rights holders to implement and impractical for developers to track, and most importantly, it does not provide for any fair compensation for the mass commercial use of works, which undermines the economic foundations of creative activity.

The results of the research confirm the proposed hypothesis: an adequate solution to the problem lies beyond the modification of existing exceptions. A paradigm shift from the logic of “exceptions” to the logic of “licensing” is necessary. In response to this need, the architecture of a hybrid model of compulsory collective licensing was proposed and substantiated. This model offers a balanced and pragmatic way forward, providing legal certainty and lawful access to data for AI developers, while at the same time guaranteeing fair remuneration for creators for the use of their intellectual property. The implementation of such a system will allow for the transformation of confrontation into cooperation, creating a sustainable ecosystem where technological innovation and human creativity can develop harmoniously. The practical implementation of this model will require political will and coordinated actions from legislators, technology companies, and creative communities at the national and international levels.

Outlining the prospects for further research requires moving from the conceptual level to practical implementation. Future work for national and international legislators should focus on three key areas. First, the development and unification of technical transparency standards. It is necessary to legally mandate that AI developers use and implement machine-readable protocols, such as C2PA, which will allow for the tracking of data provenance and automate the auditing process for CMOs. Second, the creation of infrastructure and rules for proper oversight. This includes developing accreditation criteria for CMOs, establishing clear requirements for their transparency and accountability, and forming expert commissions with all stakeholders to determine fair tariff structures. Third, international harmonization. Given the global nature of the AI market, a key task for organizations like WIPO will be to develop a framework convention that promotes mutual recognition and cooperation between national collective management systems to avoid legal arbitrage and ensure global coverage for rights holders.

References

- Abbott, R. (2020). *The reasonable robot: Artificial intelligence and the law*. Cambridge: Cambridge University Press. <https://doi.org/10.1017/9781108631761>
- Catsaros, O. (2023). Generative AI to become a \$1.3 trillion market by 2032, research finds. Retrieved from <https://www.bloomberg.com/company/press/generative-ai-to-become-a-1-3-trillion-market-by-2032-research-finds/>
- De Pauw, V., & Heremans, T. (2024). *AI and copyright: Exploring exceptions for text and data mining*. Retrieved from <https://cms-lawnow.com/en/ealerts/2024/10/ai-and-copyright-exploring-exceptions-for-text-and-data-mining>
- Eichner, S. (2023). *The copyright implications of generative AI*. Crystal City: AIPLA.
- Lucchi, N. (2025). *Generative AI and copyright: Training, creation, regulation*. Luxembourg: Publications Office of the European Union.
- Gervais, D. (2020). The machine as author. *Iowa Law Review*, 105, 2053–2106.
- Hilty, R., Hoffmann, J., & Scheuerer, S. (2020). Intellectual property justification for artificial intelligence. In J.-A. Lee, K.-C. Liu, & R. M. Hilty (Eds.), *Artificial intelligence*

- & intellectual property (pp. 1–28). Oxford: Oxford University Press.
- Kop, M. (2021). *EU artificial intelligence act: The European approach to AI*. Retrieved from <https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/>
- Lemley, M. & Casey, B. (2021). Fair learning. *Texas Law Review*, 99(4), 743-785.
- Margoni, T., & Kretschmer, M. (2022). A deeper look into the EU text and data mining exceptions: harmonisation, data ownership, and the future of technology. *GRUR International*, 71(8), 685–701. <https://doi.org/10.1093/grurint/ikac054>
- Pope, A. (2024). NYT v. OpenAI: The Times's about-face. Retrieved from <https://harvardlawreview.org/blog/2024/04/nyt-v-openai-the-timess-about-face/>
- Sag, M. (2024). Fairness and fair use. *Fordham Law Review*, 92(5), 1887-1921.
- Samuelson, P. (2023). Generative AI meets copyright. *Science*, 381(6654), 158-161. <https://doi.org/10.1126/science.adi0656>
- United States District Court for the Northern District of California. (2024). Andersen et al v. Stability AI Ltd. et al. Retrieved from <https://law.justia.com/cases/federal/district-courts/california/candce/3:2023cv00201/407208/166/>
- United States District Court for the Southern District of New York. (2024a). Authors Guild, et al. v. OpenAI, Inc. Retrieved from <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2023cv08292/606655/100/>
- United States District Court for the Southern District of New York. (2024b). The New York Times Company v. Microsoft Corporation et al. Retrieved from <https://law.justia.com/cases/federal/district-courts/new-york/nysdce/1:2023cv11195/612697/344/>
- U.S. Copyright Office. (2023). Copyright Registration Guidance: Works containing material generated by artificial intelligence. *Federal Register*, 88(59), 16190-16194.