



COPYRIGHT PROTECTION AGAINST THE USE OF WORKS AS ARTIFICIAL INTELLIGENCE TRAINING DATA IN INDONESIA

Dewi Ayu Rahayu

Law Faculty Merdeka University of Malang, Malang, Indonesia

dewi.rahayu@unmer.ac.id

Ariyanti

Law Faculty Merdeka University of Malang, Malang, Indonesia

ariyanti@unmer.ac.id

✉ corresponding email only: dewi.rahayu@unmer.ac.id

Abstract

The advancement of generative artificial intelligence has created new legal challenges in the domain of intellectual property rights, particularly copyright law. The training of AI models requires the use of massive amounts of copyrighted works as training data; however, Indonesian Law Number 28 of 2014 on Copyright does not explicitly regulate the legality of such practice. This research aims to analyze the legal standing of the use of copyrighted works as AI training data under the Indonesian copyright law, to identify existing normative gaps, and to formulate an ideal licensing model for AI development in Indonesia. Employing normative legal research through statute, conceptual, and comparative approaches, this research examines the regulatory frameworks of the United States, the European Union, Japan, and the United Kingdom as comparative material. The findings indicate that the use of copyrighted works as AI training data potentially infringes the creator's exclusive rights, particularly the right of reproduction, while no exception or limitation in Indonesia's Copyright Law specifically accommodates such use. This research proposes an AI training-data licensing model built on a collective-licensing mechanism,

accompanied by a national opt-out scheme and dataset-transparency obligations, as a normative solution that balances copyright protection with the need for artificial intelligence innovation in Indonesia.

Keywords

Copyright; AI Training Data; Generative Artificial Intelligence; Collective Licensing; Text and Data Mining

Introduction

The artificial intelligence (AI) revolution has fundamentally transformed the landscape of intellectual creation. Generative AI systems such as ChatGPT, Gemini, Claude, Midjourney, Stable Diffusion, and Meta AI are now able to produce text, images, music, and even software code with a quality that resembles human creativity. This capability does not arise from a vacuum; it stems from a machine-learning process that requires the consumption of data on a scale unprecedented in the history of information technology.

From the perspective of copyright law, the AI training process brings together three interests that remain in unresolved tension: the interest of creators who seek remuneration for every exploitation of their works; the interest of AI developers who require the broadest possible access to data in order to produce high-quality models; and the public interest in the rapid advancement of AI technology for the benefit of civilization. The tension among these three interests constitutes the core problem addressed by this research, namely the need for a legal norm capable of accommodating all three in a balanced manner.

The current legal condition, however, reflects a reality far removed from this need. Law Number 28 of 2014 on Copyright (hereinafter referred to as the 2014 Copyright Law) was drafted before the massive expansion of generative AI. The law regulates the creator's exclusive rights, which include the rights of reproduction, distribution, and the making of derivative works, yet it makes no mention whatsoever of circumstances in which copyrighted works are used as raw material for training AI systems. The limitations and exceptions to copyright set out in Articles 43 through 51 of the 2014 Copyright Law likewise do not cover use for machine-learning purposes, resulting in a significant normative gap.

This situation is not merely theoretical. Between 2022 and 2025, a wave of litigation swept across multiple jurisdictions, reflecting the magnitude of the problem. Getty Images filed suit against Stability AI

over the alleged unauthorized use of millions of copyrighted photographs as training data. The New York Times sued OpenAI and Microsoft over the use of thousands of journalistic articles. Several book publishers jointly sued Meta over the alleged unlawful use of copyrighted book collections.¹

At the international level, various jurisdictions have responded to this problem with diverse approaches. The European Union, through Directive (EU) 2019/790 on Copyright in the Digital Single Market, introduced a Text and Data Mining (TDM) exception that permits the use of works for data analysis, accompanied by an opt-out mechanism for rightsholders.² Japan, through Article 30-4 of its Copyright Act as amended in 2018, provides an even broader exception, permitting the use of works for “information analysis,” including AI training, for both commercial and non-commercial purposes.³

The United States, meanwhile, continues to rely on the fair-use doctrine developed through case-by-case judicial decisions, while the United Kingdom is undergoing a reform process to accommodate the use of works in AI training through a commercial data-mining exception mechanism. Unlike these countries, Indonesia has yet to establish a legal framework that specifically addresses this issue. The state of the art shows that academic study of copyright and AI in Indonesia has only begun to develop in recent years. A number of studies have examined the question of copyright ownership over AI-generated works,⁴ but none has specifically addressed the use of copyrighted works as AI training data from the perspective of the Indonesian legal system while simultaneously offering a concrete normative model. This is the research gap that the present study seeks to fill, not merely a gap at the level of positive norms (normative gap), but also a gap at the conceptual level (conceptual gap, namely the absence of a theoretical framework that specifically situates AI training

¹ See *Getty Images (US), Inc. v. Stability AI, Ltd.*, No. 1:23-cv-00135-UNA (D. Del. Feb. 3, 2023); *The New York Times Co. v. Microsoft Corp.*, No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).

² Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, OJ L 130, 17.5.2019, pp. 92–119, Arts. 3–4.

³ Japan Copyright Act, Article 30-4 (amended 2018), effective 1 January 2019. See Agency for Cultural Affairs (Japan), ‘General Understanding on AI and Copyright in Japan’ (March 2024).

⁴ Among others: Ari Juliano Gema, ‘Masalah Penggunaan Ciptaan Sebagai Data Masukan Dalam Pengembangan Artificial Intelligence Di Indonesia,’ *Technology and E-Commerce Law Journal*, Vol. 1, No. 1 (2023).

within the taxonomy of Indonesian copyright law), a gap at the institutional level (institutional gap, namely the unreadiness of Collective Management Organizations to handle large-scale data licensing), and a gap at the level of comparative law (comparative gap, namely the absence of a study synthesizing various international TDM models into the context of Indonesia's civil-law system).

The urgency of this research is further reinforced by the fact that Indonesia is currently in the process of developing its national AI ecosystem. The Government of Indonesia has issued the National Strategy for Artificial Intelligence 2020-2045 and various policies related to digital transformation. However, without legal certainty in the field of copyright, domestic AI developers will face substantial legal risk, while Indonesian creators lack any mechanism to protect their works from AI use without compensation.

The novelty of this research lies in the proposal of a concrete normative model in the form of an AI training-data licensing system designed to fit the characteristics of the Indonesian legal system, encompassing a collective-licensing mechanism, a national opt-out scheme, and dataset-transparency obligations for AI developers. This model differs from approaches adopted in other countries because it is tailored to the legal and institutional copyright infrastructure that already exists in Indonesia, particularly through the role of Collective Management Organizations (Lembaga Manajemen Kolektif/LMK).

Based on the foregoing, this research formulates two central problems: (1) How is the use of copyrighted works as Artificial Intelligence training data regulated from the perspective of Indonesian copyright law? and (2) What would constitute an ideal licensing concept for the use of copyrighted works as Artificial Intelligence training data so as to provide legal certainty in Indonesia? The objectives of this research are to analyze the legal standing of such practice, to identify normative gaps, to examine comparative regulation, and to formulate a licensing model suited to Indonesia's needs.

Research Method

This research constitutes normative legal research, examining legal issues on the basis of primary and secondary legal sources.⁵ The choice of a normative method is grounded in the character of the research problem, which focuses on analyzing the gap between the legal norms currently in force and the legal norms that ought to exist in

⁵ Peter Mahmud Marzuki, *Penelitian Hukum*, Revised Edition (Jakarta: Kencana Prenadamedia Group, 2016), pp. 55–56.

addressing the phenomenon of using copyrighted works as AI training data.

This research employs three approaches synergistically. First, the statute approach is used to analyze the provisions of the 2014 Copyright Law, particularly the norms concerning the creator's exclusive rights (Articles 8-10), copyright limitations and exceptions (Articles 43-51), and licensing provisions (Articles 79-92). Second, the conceptual approach is used to examine relevant legal doctrines, including exclusive-rights theory, interest-balancing theory, law-and-economics analysis, and Nonet and Selznick's theory of responsive law.⁶

Third, the comparative approach is used to examine four jurisdictions with more advanced regulatory frameworks: the United States with its fair-use doctrine, the European Union with the TDM exception under Directive 2019/790, Japan with the information-analysis exception under Article 30-4 of its Copyright Act, and the United Kingdom, which is undergoing a legislative reform process in this field.⁷

The legal materials used in this research consist of primary legal materials, comprising Indonesian statutory instruments (the 2014 Copyright Law and its implementing regulations), international instruments (the Berne Convention and the WIPO Copyright Treaty/WCT), and the legislation of comparator countries. Secondary legal materials comprise scholarly literature indexed in Scopus and accredited journals, national and international law journals, and expert opinion. Tertiary legal materials comprise legal dictionaries and encyclopedias.

Legal materials were collected through systematic library research. The analysis was conducted qualitatively through legal description, interpretation, systematization, construction, and evaluation of the norms under study.⁸ The results of this analysis were then used to construct a new normative framework in the form of the

⁶ Philippe Nonet and Philip Selznick, *Law and Society in Transition: Toward Responsive Law* (New York: Routledge, 2017), pp. 73–113.

⁷ The selection of these four jurisdictions is based on the consideration that each represents a distinct regulatory model: a common-law model relying on case-by-case tests (the United States), a model based on explicit regulation with an opt-out mechanism (the European Union), a model based on a broad exception without opt-out (Japan), and a common-law model currently transitioning toward a new commercial exception (the United Kingdom), thereby providing a comprehensive comparative spectrum for formulating the Indonesian model.

⁸ Johnny Ibrahim, *Teori dan Metodologi Penelitian Hukum Normatif* (Malang: Bayumedia Publishing, 2007), pp. 306–307.

proposed AI training-data licensing model, offered as a policy recommendation.

Result And Discussion

A. The Legal Standing of Using Copyrighted Works as AI Training Data under the Indonesian Copyright System

To understand the legal issue at stake, it is first necessary to grasp, in outline, the technical mechanism of AI training. Generative AI systems are trained through a process known as machine learning, in which an algorithm analyzes an extremely large dataset in order to identify patterns, relationships, and structures within the data. Several technical operations relevant to copyright norms occur in this process: the downloading of works, temporary copying into computer memory, tokenization of text, and the computation of vector representations of works.⁹

The legal issue does not lie in the AI's output alone, but precisely in the training stage itself. At this stage, copyrighted works are reproduced digitally on a massive scale. Although the ultimate purpose is not to present a copy of the work to end-users, the technical act of copying nonetheless occurs. Under copyright law, reproduction without the rightsholder's authorization constitutes an infringement of the creator's exclusive right.

The 2014 Copyright Law grants creators two categories of exclusive rights: moral rights and economic rights. The economic right most relevant to the issue of AI training data is the right of reproduction, as regulated under Article 9(1)(a) of the 2014 Copyright Law, which grants the creator the exclusive right to reproduce a work in any form.¹⁰

Downloading and storing a dataset containing copyrighted works for AI-training purposes textually falls within the meaning of "reproduction of a work." The phrase "in any form" in Article 9(1)(a) indicates that the legislature adopted a broad formulation, such that digital reproduction is also encompassed. This is consistent with the

⁹ Andres Guadamuz, 'The Artificial Intelligence Act and Copyright,' *WIPO Magazine* (2024); see also Ryan Abbott and Elizabeth Rothman, 'Everything is Obvious,' *UCLA Law Review*, Vol. 66, No. 1 (2019): 2–68.

¹⁰ Indonesia, Law Number 28 of 2014 on Copyright, State Gazette of the Republic of Indonesia Year 2014 Number 266, Supplement to the State Gazette Number 5599, Article 9(1)(a).

General Elucidation of the 2014 Copyright Law, which affirms the need for copyright protection in the digital era.¹¹

Beyond the right of reproduction, a number of scholars have also identified the potential infringement of the right of adaptation (Article 9(1)(c) of the 2014 Copyright Law) where an AI model subsequently produces a derivative work substantially resembling a work used in training.¹² This is particularly relevant in the context of AI models that generate images or text resembling the style or content of a specific work forming part of its training dataset.

The next step of analysis is to examine whether the 2014 Copyright Law contains any limitation or exception that could serve as a legal basis justifying the use of works as AI training data. Articles 43 through 51 of the 2014 Copyright Law set out a series of copyright limitations and exceptions, among others for the purposes of education, research, religion, security, and government.

These limitations, however, contain conditions that are not automatically satisfied in the context of commercial AI training. Article 44(1) of the 2014 Copyright Law permits the use, retrieval, reproduction, and/or alteration of a work and/or a related-rights product, in whole or in a substantial part, provided that such use does not prejudice the reasonable interests of the creator or copyright holder.¹³ The requirement of “not prejudicing reasonable interests” is a formulation adapting the Berne Convention’s three-step test, which, in the context of massive-scale use for commercial purposes, is potentially unmet.

Furthermore, research by Ari Juliano Gema concludes that copyrighted works may be used as AI input data without authorization only under three limited conditions: first, where the term of copyright protection has expired; second, where the work is released under an open license; and third, where the use falls within a limitation or

¹¹ 2014 Copyright Law, General Elucidation. See also Ari Juliano Gema, ‘Masalah Penggunaan Ciptaan Sebagai Data Masukan dalam Pengembangan Artificial Intelligence di Indonesia,’ *Technology and E-Commerce Law Journal*, Vol. 1, No. 1 (2023).

¹² Matthew Sag, ‘Copyright and Copy-Reliant Technology,’ *Northwestern University Law Review*, Vol. 103, No. 4 (2009): 1607–1681; see also Alaina Wiens, ‘Generative AI and Fair Use: A Primer for the Digital Age,’ *Harvard Journal on Legislation*, Vol. 61, No. 1 (2024): 113–178.

¹³ 2014 Copyright Law, Article 44(1). It should be noted that this article applies only to works that “have been publicly announced” and is subject to a number of cumulative conditions.

exception recognized under the 2014 Copyright Law.¹⁴ For massive-scale commercial uses such as the training of large AI models (Large Language Models/LLMs), none of these three conditions is inherently satisfied.

It may therefore be concluded, from the analysis of the 2014 Copyright Law, that a genuine normative gap exists: there is no provision that expressly permits the use of copyrighted works as AI training data, yet equally no norm that expressly prohibits it with a clear enforcement procedure. It is precisely this legal uncertainty that must be resolved through normative reconstruction or legislative reform.

B. Comparative Study of the Regulation of Using Works as AI Training Data

The United States addresses this issue through the fair-use doctrine set out in 17 U.S.C. § 107. The fair-use test comprises four factors: (1) the purpose and character of the use, including whether it is commercial in nature; (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 the work.¹⁵

In the AI context, the argument most frequently advanced by AI developers is that training constitutes transformative use that is, use of a work for a purpose fundamentally different from its original purpose. This argument draws on the U.S. Court of Appeals' decision in *Authors Guild v. Google, Inc.*, in which the digitization of books for search-engine purposes was held to constitute fair use on account of its transformative character.¹⁶

However, the application of this principle in the context of AI training remains contested. The U.S. Copyright Office report published in May 2025 (Copyright and Artificial Intelligence Part 3: Generative AI Training) concluded that fair use does not automatically apply to the entirety of AI training activity, and that copying works on a massive scale for commercial purposes may fail to satisfy the fair-use test, particularly under the fourth factor (market effect).¹⁷ A countervailing

¹⁴ Ari Juliano Gema, 'Masalah Penggunaan Ciptaan Sebagai Data Masukan Dalam Pengembangan Artificial Intelligence Di Indonesia,' *Technology and E-Commerce Law Journal*, Vol. 1, No. 1 (2023).

¹⁵ 17 U.S.C. § 107 (2018). This fair-use doctrine is a continuation of common-law principles codified since the Copyright Act of 1976.

¹⁶ *Authors Guild v. Google, Inc.*, 804 F.3d 202 (2d Cir. 2015), *cert. denied*, 578 U.S. 945 (2016). See the analysis in Paul Goldstein, *Copyright: Principles, Law and Practice*, 3rd ed. (New York: Aspen Publishers, 2020).

¹⁷ U.S. Copyright Office, 'Copyright and Artificial Intelligence Part 3: Generative AI Training' (May 2025). The report also concludes that the development of scalable licensing mechanisms is greatly needed.

argument, which deserves balanced consideration, is that an overly restrictive application of fair use risks impeding AI research and innovation that carries substantial social benefit, including the development of language models for low-resource languages.

That report identifies four licensing options that might be developed: voluntary direct licensing, voluntary collective licensing, extended collective licensing (ECL), and compulsory licensing.¹⁸ Although the report does not explicitly recommend any particular option, it encourages the development of scalable licensing mechanisms and the growth of a voluntary market before any legislative intervention is undertaken. A further relevant development is the settlement of the class-action lawsuit *Bartz et al. v. Anthropic PBC* in the United States, which suggests that resolution through collective, settlement-based compensation schemes is beginning to emerge as a practical precedent, even though it has not yet become a generally binding legal norm.¹⁹

The European Union, by contrast, has adopted a more structured approach through Directive (EU) 2019/790 on Copyright in the Digital Single Market (the Copyright Directive/CDSMD). This Directive introduces two TDM exceptions: Article 3, providing a mandatory exception for research organizations and cultural heritage institutions, and Article 4, providing an exception for commercial use provided that the rightsholder has not objected (opted out) through machine-readable means.²⁰

As it has developed, the EU Artificial Intelligence Act (Regulation 2024/1689), in force since August 2024, has strengthened the linkage between AI developers' obligations and the copyright framework. Article 53 of the AI Act requires providers of general-purpose AI (GPAI) models to implement a copyright-compliance policy, to respect the TDM opt-out mechanism under Article 4(3) of the CDSMD, and to publish a sufficiently detailed summary of the training data used.²¹ This transparency obligation, as examined by Buick, is fundamentally a necessary but not, by itself, sufficient instrument for resolving the question of creator compensation; its effectiveness depends heavily on the strength of the underlying copyright norms,

¹⁸ U.S. Copyright Office, 'Copyright and Artificial Intelligence Part 3: Generative AI Training,' Section V on Licensing for AI Training (May 2025).

¹⁹ See *Settlement Agreement* (Class Action Settlement Agreement), *Bartz et al. v. Anthropic PBC* (N.D. Cal. Sept. 5, 2025).

²⁰ Directive (EU) 2019/790, Arts. 3–4. See the in-depth analysis in Martin Senftleben et al., 'The Recommendation on Measures to Safeguard Freedom of Expression and Information in Copyright Law,' *European Intellectual Property Review*, Vol. 41, No. 4 (2019): 208–228.

²¹ Regulation (EU) 2024/1689 (AI Act), Art. 53(1)(c)-(d).

such that dataset transparency needs to be accompanied by other substantive mechanisms, such as collective licensing or remuneration obligations.²²

The first significant piece of European case law concerning the use of TDM for AI training is the decision of the Hamburg Regional Court, Germany, in *Kneschke v. LAION*, delivered in September 2024. The court held that downloading copyrighted photographs to build an AI training dataset fell within the TDM exception for scientific research under the German implementation of Article 3 of the CDSMD (§ 60d UrhG).²³

The EU model offers a relatively balanced framework: use for research is permitted unconditionally (subject to lawful access), while commercial use is permitted but subject to an opt-out mechanism enabling rightsholders to exclude their works. Nevertheless, various researchers and stakeholders have noted that the opt-out mechanism continues to face substantial technical and implementation challenges in practice.²⁴

Japan represents the most pro-innovation jurisdiction on this issue. Through the 2018 amendment to its Copyright Act, effective 1 January 2019, Article 30-4 permits the use of copyrighted works for the purpose of “information analysis,” defined as the extraction and analysis of language, sound, images, and other information from a large volume of information, including comparison, classification, and other forms of analysis.²⁵

This exception has several important characteristics distinguishing it from the EU’s TDM exception. First, Japan’s exception applies to all entities, whether commercial or non-commercial. Second, there is no opt-out mechanism enabling rightsholders to exclude their works. Third, the exception does not explicitly require that a work be

²² Adam Buick, ‘Copyright and AI Training Data—Transparency to the Rescue?’, *Journal of Intellectual Property Law & Practice*, Vol. 20, No. 3 (2025): 182–192.

²³ Regional Court of Hamburg, *Decision of 27 September 2024 (Kneschke v. LAION)*. Reviewed in Philipp Homar and Philipp Pointner, ‘Landmark Court Decision in the EU: Copyright Permissibility of Text and Data Mining for the Purpose of AI Training,’ *RAILS Blog* (ai-laws.org), 5 January 2025.

²⁴ European Parliament, ‘AI and Copyright: The Training of General-Purpose AI,’ Briefing PE 769.585 (2025); see also João Quintais et al., ‘Generative AI and Copyright: A Fair Use Analysis,’ *European Intellectual Property Review*, Vol. 46, No. 3 (2024): 147–163.

²⁵ Japan Copyright Act, Art. 30-4, as amended by the Act for Partial Amendment of the Copyright Act (Act No. 30 of 2018). See Tatsuhiro Ueno, ‘Data Mining Exception and Other Copyright Issues in Japan,’ *Journal of the Copyright Society of the U.S.A.*, Vol. 67, No. 3 (2020): 389–420.

lawfully obtained. The exception is inapplicable only where the purpose of use is to “enjoy the expression” of a work, or where the use “unreasonably prejudices the interests of the copyright holder.”²⁶

Japan’s highly liberal approach has begun to face pressure. In December 2024, the Japan Newspaper Publishers & Editors Association (Nihon Shinbun Kyokai) formally requested that the government review this framework.²⁷ In early 2025, the Japanese government began signaling the possibility of amending Article 30-4, or adding new provisions, to provide more adequate protection for rightsholders, particularly regarding dataset-transparency obligations and the use of content obtained from unauthorized sources.

The United Kingdom, by contrast, applies the doctrine of fair dealing, which is narrower than U.S. fair use. Following Brexit, the UK has developed its own legal framework. After a series of consultations since 2021, in December 2024 the UK Government launched a consultation on a new exception for commercial data mining that would allow AI developers to use copyrighted works on a broad scale, subject to an opt-out mechanism for rightsholders and transparency obligations.²⁸

Five key lessons emerge from this comparative study. First, no single model is universally regarded as ideal; each jurisdiction balances the interests of creators and AI developers differently, in accordance with its own legal tradition and creative-industry policy. Second, explicit legal certainty (as in the EU and Japanese models) provides a clearer investment climate than a case-by-case approach (as in the U.S. model). Third, dataset-transparency obligations have become an almost uniform regulatory trend across advanced jurisdictions, even though, as the EU experience demonstrates, transparency alone does not resolve the question of creators’ economic compensation. Fourth, collective-licensing mechanisms have emerged as the most promising solution for accommodating massive-scale training-data use without eliminating creators’ right to compensation. Fifth, the opt-out principle (whereby rightsholders exclude their works from AI training use) is conceptually better suited to fostering AI innovation than the opt-in principle

²⁶ Agency for Cultural Affairs (Japan), ‘General Understanding on AI and Copyright in Japan’ (March 2024).

²⁷ Keisen Associates, ‘Japan Signals Potential Copyright Law Revisions in Response to Generative AI’ (March 2025).

²⁸ UK Government, ‘Copyright and AI: Consultation on a New Commercial Exception for Text and Data Mining’ (December 2024). See also Julia Reda and Paul Keller, ‘Exceptions for Text and Data Mining in the Digital Single Market Directive,’ *International Review of Intellectual Property and Competition Law*, Vol. 50, No. 7 (2019): 844–856.

(whereby rightsholders must actively grant permission), although the opt-out mechanism must be designed so as not to impose a disproportionate burden on individual creators as illustrated by the EU's continuing difficulties in implementing opt-out through machine-readable means.

C. Constructing an AI Training-Data Licensing Model for Indonesia

The construction of this licensing model rests on four theoretical pillars. First, Exclusive Rights Theory, which affirms that creators hold an inherent right over every exploitation of their works, including for AI-training purposes.²⁹ Second, Balance Theory, as advocated by scholars such as William Fisher and Lawrence Lessig, which emphasizes that copyright must balance the interests of creators against the interests of innovation and public access.³⁰

Third, the Economic Analysis of Law developed by Posner, which analyzes the efficiency of various licensing models in internalizing the positive externalities of AI innovation while providing adequate incentives for creators.³¹ Fourth, Nonet and Selznick's theory of responsive law, which provides a framework for designing regulation capable of adapting to technological change without requiring statutory revision each time a new development occurs.³²

Building on the comparative study and the theoretical framework outlined above, this research formulates an AI training-data licensing model for Indonesia comprising five interrelated core elements: a). An AI Training Data License, a sui generis legal instrument distinct from conventional copyright licenses, specifically governing the use of copyrighted works for AI training purposes. This license is designed not merely for a one-to-one relationship between creator and AI developer, but for operation at massive scale through an institutional mechanism; b). A Collective Licensing mechanism operated through Collective Management Organizations (LMK), which already exist within the Indonesian legal system under Articles 87-93 of the 2014 Copyright Law. An LMK acts as an intermediary that aggregates rights from creators and grants authorization to AI developers, while also

²⁹ Justin Hughes, 'The Philosophy of Intellectual Property,' *Georgetown Law Journal*, Vol. 77, No. 2 (1988): 287–366.

³⁰ Lawrence Lessig, *The Future of Ideas: The Fate of the Commons in a Connected World* (New York: Random House, 2001); William Fisher, *Promises to Keep: Technology, Law, and the Future of Entertainment* (Stanford: Stanford University Press, 2004), pp. 199–258.

³¹ Richard A. Posner, *Economic Analysis of Law*, 9th ed. (New York: Wolters Kluwer Law & Business, 2014), pp. 53–72.

³² Nonet and Selznick, *Law and Society in Transition*, pp. 73–113.

administering royalty payments. This mechanism enables AI developers to obtain access to datasets at the scale required without having to negotiate authorization with each creator individually;³³ c). A National Opt-Out Scheme granting rightsholders the right to declare that their works may not be used as AI training data, either individually or collectively through an LMK. The opt-out mechanism should be designed to be technically simple for instance, through the addition of standardized metadata or registration in a national registry administered by the Directorate General of Intellectual Property so as not to burden creators who lack technical expertise;³⁴ d). A Dataset-Transparency Obligation requiring AI developers to adequately disclose the source, categories, and scope of the dataset used in the training process. This obligation gives creators a basis for ascertaining whether their works have been used, while also providing the basis for calculating proportionate royalties. As affirmed in Section B above, dataset transparency should be understood as an administrative prerequisite rather than a stand-alone substantive solution; it becomes effective only when integrated with a collective-licensing mechanism and a binding remuneration scheme;³⁵ e). A Normative Reconstruction of the 2014 Copyright Law through the addition of provisions that expressly define the use of works for machine learning, regulate the AI training-data licensing mechanism, establish proportionate royalty standards, and impose sanctions for breach of the dataset-transparency obligation. This reconstruction could be achieved through the addition of new articles in a dedicated chapter, or through a government regulation implementing the 2014 Copyright Law.

The proposed licensing model requires at least three implementation mechanisms to be developed simultaneously. First, a rights-registration mechanism (rights registry) enabling rightsholders to register their works together with their opt-out or opt-in preference. This registry could be administered by the Directorate General of

³³ 2014 Copyright Law, Articles 87-93 on Collective Management Organizations. On comparable institutional models of collective licensing in other jurisdictions, see Daniel Gervais, 'Collective Management of Copyright: Theory and Practice in the Digital Age,' in Daniel Gervais (ed.), *Collective Management of Copyright and Related Rights*, 3rd ed. (Alphen aan den Rijn: Kluwer Law International, 2016), ch. 1.

³⁴ A similar mechanism is proposed in Martin Senftleben, 'Generative AI and Author Remuneration,' *International Review of Intellectual Property and Competition Law*, Vol. 54, No. 8 (2023): 1049–1083.

³⁵ The dataset-transparency obligation is consistent with Art. 53(1)(d) of the EU AI Act. See also Nathalie A. Smuha et al., 'How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission's Proposal for an Artificial Intelligence Act' (SSRN, 2021).

Intellectual Property in cooperation with existing LMKs, and should be built in a digitally accessible format.

Second, a fair and transparent royalty calculation and distribution mechanism. Royalty calculation should take into account several factors: the type of work used, the volume of use within the dataset, and the work's contribution to the capability of the resulting AI model. Royalty distribution could be carried out proportionately, based on the number of works registered within the dataset.³⁶

Third, an effective enforcement mechanism. A sound licensing model must be accompanied by an enforcement mechanism capable of imposing real sanctions on AI developers that use copyrighted works without satisfying licensing requirements. Given the cross-border nature of the AI ecosystem, enforcement requires international cooperation and efficient alternative dispute-resolution mechanisms.³⁷

One advantage of the proposed model is its compatibility with the legal infrastructure that already exists in Indonesia. The 2014 Copyright Law has already provided the basis for the existence of LMKs, and several LMKs currently operate in Indonesia, such as YKCI (Yayasan Karya Cipta Indonesia), LMKN (Lembaga Manajemen Kolektif Nasional), KCI, and WAMI. The AI training-data licensing model can be developed upon this existing institutional foundation, without requiring the establishment of an entirely new institution.³⁸

Nevertheless, implementing this model also requires strengthening LMK capacity, particularly regarding the technical ability to verify the use of works within AI datasets, to track royalty distribution, and to build a reliable digital registry. This capacity-building will likely require significant investment and technical cooperation with parties possessing information-technology expertise.³⁹

³⁶ On royalty-distribution systems in the digital context, see Daniel Gervais and João Pedro Quintais (eds.), *Collective Management of Copyright and Related Rights*, 4th ed. (Alphen aan den Rijn: Kluwer Law International, 2025).

³⁷ On copyright enforcement in the cross-border AI context, see Jane Ginsburg, 'People Not Machines: Authorship and What It Means in the Berne Convention,' *International Review of Intellectual Property and Competition Law*, Vol. 49, No. 2 (2018): 131–135.

³⁸ On LMKs in Indonesia, see Budi Agus Riswandi, *Hak Cipta di Internet: Aspek Hukum dan Permasalahannya di Indonesia* (Yogyakarta: FH UII Press, 2009), pp. 112–128. On the role of LMKs in the digital era, see also Agus Sardjono, *Hak Kekayaan Intelektual dan Pengetahuan Tradisional* (Bandung: Alumni, 2010).

³⁹ On strengthening LMK capacity in the digital context, see WIPO, *Guide on Managing Intellectual Property for the Information and Communications Technology Sector* (Geneva: WIPO, 2016), pp. 87–102.

The proposed model is consistent with evolving international regulatory trends. Both the European Union, through the AI Act, and the United States, through the 2025 Copyright Office report, underscore the need for scalable collective-licensing mechanisms and dataset transparency. The Indonesian model offers a synthesis of these approaches, tailored to Indonesia's specific conditions: a civil-law system, existing LMK infrastructure, and the need to both support and regulate a rapidly developing AI ecosystem.⁴⁰

At the international level, the World Intellectual Property Organization (WIPO) is currently discussing the possible development of an international instrument on AI and copyright within the framework of the Standing Committee on Copyright and Related Rights (SCCR). Indonesia's active participation in this process would allow its national interests as a country with a rich creative ecosystem to be reflected in the international norms that eventually emerge.⁴¹

D. Implementation Challenges and a Policy Roadmap

Notwithstanding the conceptual strength of the proposed model, this research also identifies a number of implementation challenges that must be anticipated from the outset, given that empirical experience in other jurisdictions shows that the gap between sound policy design on paper and its actual implementation is frequently the source of regulatory failure.

The first challenge is the institutional-capacity gap. LMKs in Indonesia currently operate primarily in the management of royalties for musical works and songs, at a scale of transaction and coverage far smaller than what is required to manage AI datasets, which may encompass millions to billions of individual works across multiple categories (text, image, audio, video). Without investment in a capable rights-management information system, LMKs risk being unable to perform verification and royalty-distribution functions accurately, such that a normatively ideal licensing model could nonetheless fail operationally.

The second challenge is the risk of fragmented creator participation. EU experience shows that an individual, machine-readable opt-out mechanism faces a low rate of compliance from AI

⁴⁰ A similar synthesis is proposed for developing countries in Carolyn Deere Birkbeck, *The Political Economy of Intellectual Property Rights and Development* (Geneva: ICTSD, 2016); and Jyh-An Lee, Reto M. Hilty, and Kung-Chung Liu (eds.), *Artificial Intelligence and Intellectual Property* (Oxford: Oxford University Press, 2021).

⁴¹ WIPO, 'Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence' (2020), WIPO/IP/AI/2/GE/20/1 REV; WIPO, *Technology Trends 2019: Artificial Intelligence* (Geneva: WIPO, 2019).

developers, particularly those operating outside the EU's jurisdiction.⁴² This condition could recur in Indonesia if the national opt-out scheme is not accompanied by a legal obligation that is binding extraterritorially, or at least binding upon AI service providers that operate commercially within Indonesia.

The third challenge is the jurisdictional limitation in reaching cross-border AI developers. Most large-scale generative AI models are developed by companies based outside Indonesia. The effectiveness of licensing and dataset-transparency obligations depends heavily on Indonesia's ability to enforce such norms against foreign entities, whether through bilateral agreements, cooperation within multilateral forums such as WIPO's SCCR, or an extraterritorial approach similar to that of the EU AI Act, which reaches GPAI providers that place their products on the EU market regardless of where those products were developed.

The fourth challenge is the need to balance creator protection against the investment climate for domestic AI developers, which remain in an early stage of growth. Overly burdensome licensing and royalty obligations risk driving domestic AI developers to relocate their operations to jurisdictions with a more permissive copyright regime, as illustrated by Japan's precedent of deliberately designing a broad exception to attract AI research investment.

Based on the identification of these challenges, this research proposes a three-stage implementation roadmap. The first stage (short term, 1-2 years) focuses on issuing a government regulation implementing the 2014 Copyright Law that expressly defines the legal status of using works for AI training, together with minimum dataset-transparency obligations. The second stage (medium term, 2-4 years) focuses on strengthening LMK institutional capacity and building a digitally integrated national rights registry. The third stage (long term, beyond 4 years) focuses on harmonization with evolving international norms through WIPO's SCCR forum, together with a comprehensive evaluation of the effectiveness of the collective-licensing and national opt-out schemes based on real implementation data, so as to open the way for continuous refinement of the model.

Conclusion

Based on the analysis undertaken, this research draws two principal conclusions that answer the research questions posed. First, the use of copyrighted works as Artificial Intelligence training data,

⁴² European Parliament, 'AI and Copyright: The Training of General-Purpose AI,' Briefing PE 769.585 (2025).

viewed from the perspective of Indonesian copyright law, potentially infringes the creator's exclusive rights, particularly the right of reproduction guaranteed under Article 9(1)(a) of the 2014 Copyright Law. The 2014 Copyright Law contains no limitation or exception that explicitly accommodates the use of works for machine learning. The absence of an explicit norm creates a condition of legal ambiguity that risks harming both parties: creators, who lack any mechanism to claim compensation, and AI developers, who operate under legal uncertainty. There exist a normative gap, a conceptual gap, an institutional gap, and a comparative gap that must be resolved through normative reconstruction or legislative reform.

Second, the ideal licensing concept for the use of copyrighted works as AI training data, so as to provide legal certainty in Indonesia, is a model based on a collective-licensing mechanism operated through the Collective Management Organizations that already exist within Indonesia's copyright ecosystem, complemented by three core elements: (1) a national opt-out scheme enabling rightsholders to exclude their works from AI training use through a simple and affordable mechanism; (2) a dataset-transparency obligation requiring AI developers to disclose the source and scope of their training data; and (3) a proportionate and transparent royalty-distribution mechanism. This model balances creators' interest in fair compensation with the need for AI innovation, while providing the legal certainty required by all stakeholders — with the caveat that its successful implementation depends heavily on strengthening LMK institutional capacity and on Indonesia's ability to reach cross-border AI developers.

Recommendations

Based on the conclusions of this research, a concerted effort by all stakeholders is required to build a legal framework capable of accommodating the development of artificial intelligence technology without disregarding copyright protection. Regulatory strengthening should be directed toward providing legal certainty regarding the use of copyrighted works as AI training data, establishing an effective and equitable licensing mechanism, and applying principles of transparency and accountability in AI developers' use of datasets. In addition, institutional strengthening is required, together with the development of systems supporting copyright management in the digital era, including a mechanism for recording rightsholders' preferences and for the proportionate distribution of economic benefits. The Government, through the Directorate General of Intellectual Property, is advised to prioritize the drafting of a government regulation as a short-term step, as outlined in the policy roadmap in Section III.D, while simultaneously

strengthening LMKs' technical capacity and actively participating in the discussion of international norms within WIPO's SCCR forum. Further research is also needed to examine both the technical and juridical implementation of the proposed licensing model, including empirical study of LMKs' institutional readiness, so as to produce a regulatory system that is adaptive, effective, and balanced in protecting the interests of copyright holders, fostering technological innovation, and serving the public interest.

References

- Birkbeck, Carolyn Deere, *The Political Economy of Intellectual Property Rights and Development* (Geneva: ICTSD, 2016); and Jyh-An Lee, Reto M. Hilty, and Kung-Chung Liu (eds.), *Artificial Intelligence and Intellectual Property* (Oxford: Oxford University Press, 2021).
- Buick, Adam, 'Copyright and AI Training Data—Transparency to the Rescue?,' *Journal of Intellectual Property Law & Practice*, Vol. 20, No. 3 (2025): 182–192.
- Hughes, Justin, 'The Philosophy of Intellectual Property,' *Georgetown Law Journal*, Vol. 77, No. 2 (1988): 287–366.
- Gema, Ari Juliano, 'Masalah Penggunaan Ciptaan Sebagai Data Masukan Dalam Pengembangan Artificial Intelligence Di Indonesia,' *Technology and E-Commerce Law Journal*, Vol. 1, No. 1 (2023).
- Guadamuz, Andres, 'The Artificial Intelligence Act and Copyright,' *WIPO Magazine* (2024); see also Ryan Abbott and Elizabeth Rothman, 'Everything is Obvious,' *UCLA Law Review*, Vol. 66, No. 1 (2019): 2–68.
- Gervais, Daniel, 'Collective Management of Copyright: Theory and Practice in the Digital Age,' in Daniel Gervais (ed.), *Collective Management of Copyright and Related Rights*, 3rd ed. (Alphen aan den Rijn: Kluwer Law International, 2016), ch. 1.
- Gervais, Daniel and João Pedro Quintais (eds.), *Collective Management of Copyright and Related Rights*, 4th ed. (Alphen aan den Rijn: Kluwer Law International, 2025).
- Ginsburg, Jane, 'People Not Machines: Authorship and What It Means in the Berne Convention,' *International Review of Intellectual Property and Competition Law*, Vol. 49, No. 2 (2018): 131–135.
- Goldstein, Paul, *Copyright: Principles, Law and Practice*, 3rd ed. (New York: Aspen Publishers, 2020).
- Ibrahim, Johnny, *Teori dan Metodologi Penelitian Hukum Normatif* (Malang: Bayumedia Publishing, 2007).
- Marzuki, Peter Mahmud, *Penelitian Hukum*, Revised Edition (Jakarta: Kencana Prenadamedia Group, 2016).
- Nonet, Philippe and Philip Selznick, *Law and Society in Transition: Toward Responsive Law* (New York: Routledge, 2017).

- Lessig, Lawrence, *The Future of Ideas: The Fate of the Commons in a Connected World* (New York: Random House, 2001); William Fisher, *Promises to Keep: Technology, Law, and the Future of Entertainment* (Stanford: Stanford University Press, 2004), pp. 199–258.
- Posner, Richard A., *Economic Analysis of Law*, 9th ed. (New York: Wolters Kluwer Law & Business, 2014), pp. 53–72.
- Quintais, João et al., ‘Generative AI and Copyright: A Fair Use Analysis,’ *European Intellectual Property Review*, Vol. 46, No. 3 (2024): 147–163.
- Reda, Julia and Paul Keller, ‘Exceptions for Text and Data Mining in the Digital Single Market Directive,’ *International Review of Intellectual Property and Competition Law*, Vol. 50, No. 7 (2019): 844–856.
- Riswandi, Budi Agus, *Hak Cipta di Internet: Aspek Hukum dan Permasalahannya di Indonesia* (Yogyakarta: FH UII Press, 2009), pp. 112–128. On the role of LMKs in the digital era, see also Agus Sardjono, *Hak Kekayaan Intelektual dan Pengetahuan Tradisional* (Bandung: Alumni, 2010).
- Sag, Matthew, ‘Copyright and Copy-Reliant Technology,’ *Northwestern University Law Review*, Vol. 103, No. 4 (2009): 1607–1681; see also Alaina Wiens, ‘Generative AI and Fair Use: A Primer for the Digital Age,’ *Harvard Journal on Legislation*, Vol. 61, No. 1 (2024): 113–178.
- Senftleben, Martin, ‘Generative AI and Author Remuneration,’ *International Review of Intellectual Property and Competition Law*, Vol. 54, No. 8 (2023): 1049–1083.
- Senftleben, Martin et al., ‘The Recommendation on Measures to Safeguard Freedom of Expression and Information in Copyright Law,’ *European Intellectual Property Review*, Vol. 41, No. 4 (2019): 208–228.
- Smuha, Nathalie A. et al., ‘How the EU Can Achieve Legally Trustworthy AI: A Response to the European Commission’s Proposal for an Artificial Intelligence Act’ (SSRN, 2021).
- Ueno, Tatsuhiro, ‘Data Mining Exception and Other Copyright Issues in Japan,’ *Journal of the Copyright Society of the U.S.A.*, Vol. 67, No. 3 (2020): 389–420.
- Agency for Cultural Affairs (Japan), ‘General Understanding on AI and Copyright in Japan’ (March 2024).
- Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market, OJ L 130, 17.5.2019, pp. 92–119, Arts. 3–4.
- Copyright Act of 1976.
- European Parliament, ‘AI and Copyright: The Training of General-Purpose AI,’ Briefing PE 769.585 (2025).
- Getty Images (US), Inc. v. Stability AI, Ltd., No. 1:23-cv-00135-UNA (D. Del. Feb. 3, 2023); The New York Times Co. v. Microsoft Corp., No. 1:23-cv-11195 (S.D.N.Y. Dec. 27, 2023).
- Indonesia, Law Number 28 of 2014 on Copyright, State Gazette of the Republic of Indonesia Year 2014 Number 266, Supplement to the State Gazette Number 5599, Article 9(1)(a).

- Japan Copyright Act, Article 30-4 (amended 2018), effective 1 January 2019.
See Agency for Cultural Affairs (Japan), ‘General Understanding on AI and Copyright in Japan’ (March 2024).
- Keisen Associates, ‘Japan Signals Potential Copyright Law Revisions in Response to Generative AI’ (March 2025).
- U.S. Copyright Office, ‘Copyright and Artificial Intelligence Part 3: Generative AI Training’ (May 2025). The report also concludes that the development of scalable licensing mechanisms is greatly needed.
- Regional Court of Hamburg, *Decision of 27 September 2024 (Kneschke v. LAION)*. Reviewed in Philipp Homar and Philipp Pointner, ‘Landmark Court Decision in the EU: Copyright Permissibility of Text and Data Mining for the Purpose of AI Training,’ *RAILS Blog* (ai-laws.org), 5 January 2025.
- Regulation (EU) 2024/1689 (AI Act)
- Settlement Agreement* (Class Action Settlement Agreement), *Bartz et al. v. Anthropic PBC* (N.D. Cal. Sept. 5, 2025).
- WIPO, *Guide on Managing Intellectual Property for the Information and Communications Technology Sector* (Geneva: WIPO, 2016), pp. 87–102.
- WIPO, ‘Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence’ (2020), WIPO/IP/AI/2/GE/20/1 REV; WIPO, *Technology Trends 2019: Artificial Intelligence* (Geneva: WIPO, 2019).