Research Article

Legal Ethics of Generative Artificial Intelligence Training Data Usage in Large Language Models: Copyright and Personal Data Perspective

Galih Dwi Ramadhan[1], Damar Sugeng Utomo[1*], Jihyun Park[2]
[1]Faculty of Law, Universitas Islam Indonesia, Indonesia
[2]Faculty of Law, Youngsan University, South Korea
[*]damar.sugeng@uii.ac.id

ABSTRACT

This research originates from security and privacy risks associated with Generative AI Large Language Models (LLMs), especially concerning data used for training in relation to Intellectual Property Rights (IPR) and Personal Data Protection (PDP). The study seeks to examine existing regulatory guidance and develop a national framework that protects user privacy and IPR. The method employs normative legal research using statutory and conceptual approaches, by reviewing the legal frameworks of Indonesia, South Korea, the European Union (EU), and Japan, along with direct testing of LLM prompts. The findings show that LLMs continuously absorb users' input prompts, data, and corrections to support deeper model training, thereby revealing significant security vulnerabilities under PDP laws. Moreover, in the absence of strong human creative control and predictability, each incoming data item may potentially breach privacy and copyright laws and may also damage original creators. Therefore, legal governance should move toward explicit licensing, fair compensation, and technical verification standards such as provenance labeling. The study concludes that mandatory human-centric supervision, explicit user consent, and freedom of choice are essential legal limits. Ultimately, effective AI governance depends on a shared responsibility model: software engineers must implement technical transparency, expert oversight, and responsive complaint mechanisms, while end users must select platforms lawfully, comply strictly with output licensing requirements, and actively monitor the system.

Keywords: Artificial Intelligence; Human Centricity; Intellectual Property Rights; Personal Data; Principles.

A. INTRODUCTION

The emergence of the digital age through computerization has introduced changes across multiple domains, including the development of creative works (Andini et al., 2021). Given that both the digital age and creations in digital form are unavoidable, they have become increasingly common and can no longer be separated from everyday life (Simatupang, 2021). Although technology enhances recognition of work ownership, it also raises challenges related to piracy and disputes over the economic returns produced by creative works (Mahmud & Roisah, 2023). One example of technology that is currently being widely used by humans to assist their needs known as Artificial Intelligence ("AI").

As technology advances, AI capabilities are strengthened by generating data that is similar to its training data (Tusquellas et al., 2024). AI that can

create a work is referred to Generative AI, which is a tool in the form of software that can create new content, such as text, computer code, images, audio, voice, and video, by responding the input prompts of its users. The focus of Generative AI refers to the application of algorithms used on various data stored as data training, which is then processed to produce new data.

The widespread presence of Generative AI in society today raises concerns that arise among AI software developers and, similarly, among users regarding the digital data that are transformed into training data when developing Generative AI software. The key issues with training data used in Generative AI relate to the origins of that data. These sources include a range of options, from open source and private data to information provided by AI users, which users enter through command prompts in the Generative AI-based software they use. Some of the most extensively developed and most widely used Generative AI models among most people are Large Language Models (LLM), such as the GPT series by OpenAI, Gemini by Google, the Llama series by Meta, and the Claude series by Anthropic. LLM function by employing deep learning architectures to predict the next output. Deep learning, in turn, results from the use of artificial neural networks that carry out a form of replication of human brain performance functions (Stahl et al., 2023). This advanced predictive capability enables these systems to generate contextual, relevant, and creative human-like text, effectively positioning LLM-based Generative AI models specialized in language and text outputs. Over time, LLM have evolved and can now produce diverse forms of digital content beyond language and text; for example, Grok AI developed by xAI. Given this condition, LLM are designed to analyze and process input data in the form of language and text as commands, which are then handled in accordance with the training data they have in order to generate new output data based on the instructions provided by the user.

The feature that sets LLM apart from other AI models is their capacity to process and deliver sophisticated reasoning. This ability is obtained from the processing of scientific results which become the data source of the AI LLM. Main characteristics of LLM are having very broad parameters so that they can learn and show complex language styles because the amount of training data used as a basis is very large, having the ability to perform various types of tasks from simple to complex such as analysis, resumes, and data generalization. LLM have two training stages as the foundation, namely pre-training originating from very large data clusters and fine-tuning taken from more specific data clusters. LLM has a capability called chain of thought, where the LLM will answer all tasks by predict the next output based on the prompt provided (Guo et al., 2025).

Data clusters found in LLM play a vital role as a foundation in infrastructure to be utilized as training data for LLM development. In LLM, data clusters can be divided into five categories, pre-training corpora, instruction alignment or instruction fine-tuning, learning preferences, evaluation, and traditional natural language processing ("NLP") datasets. In the management of such training data, the data clusters used to train LLM are often obtained from public data clusters that available on internet, scientific publication that have a significant impact, the most frequently cited data clusters, specific data clusters that can only be accessed with payment, systems that provide easily accessible data sets as a guide for verification, and industries that provide data clusters as a comparison data and data used in general (Liu et al., 2025). Datacenter for AI training contain and consist of thousands domain-specific accelerator ("DSA") to train their deep learning models. As for example, Zhao recommends minimum speed for Meta. Zhao finds that the cluster are rely on data storage and ingestion ("DSI") which responsible for storing exabytes of training data and serving it at ten terabytes per second. This kind of condition will lead to reality that DSI will dominate large-scale training resource because coordinated training, dataset storage and reading, and pre-processing will load in DSI infrastructure (Zhao et al., 2022).

Large language models have shown outstanding performance across many domains. Such strong performance is driven by data, which plays a central role. In the training of large language models, the most critical requirement is refined data that is submitted to and ingested by a data center. Before being stored in a database as reference data (Liang et al., 2026), this type of data must be checked. Cooper explains that, in large language models, memorization can occur either consciously or unconsciously, although we do not actually know whether it exists. This memorization can be demonstrated through the successful extraction of, or reproduction from, memorized data (regurgitation). However, in some cases, a large language model indicates that it has not memorized the data and that the data was not retained during training. This situation creates a grey area because there is no certainty about the data, particularly in generative AI systems that use prompts as the primary command. Cooper characterized this as epistemic uncertainty, which is distinct from the legal system, where there is a burden of proof, findings of fact that can be reopened, and expert analysis (Cooper & Grimmelmann, 2025). Ethical and equitable regulations must be established to ensure that AI technologies uphold fundamental human rights and core ethical principles (Budiyanti et al., 2025).

Generative AI within large language models (LLMs) involves legal issues, particularly intellectual property rights ("IPR") and personal data protection ("PDP"). The data used to train LLMs certainly

includes materials protected by IPR, and it is also possible that training involves personal information or data relating to a legal subject. Accordingly, law proceeds alongside technological progress. Training data for LLMs that is dispersed and obtained without regard for jurisdictional boundaries clearly needs clearly defined limits, so as to avoid breaching the prohibitory provisions governing IPR and PDP under a country's national law.

Intellectual property rights in Indonesia are traditionally defined as human intellectual creations in science, art, and literature, and computer software is protected under Law No. 28 of 2014 on Copyright (Ramadhan, 2022). Even though current Indonesian legal frameworks narrowly treat artificial intelligence as a computer program under Article 40(1)(s) of the Copyright Law and classify its source code as a compilation of works under Article 40(1)(p), the complexities of contemporary generative AI require a comprehensive comparative analysis. Within the IPR legal system, moral doctrines protect individual owners from violations by third parties. Therefore, IPR protection is based on an approach centered on individual rights (Roisah, 2015).

The European Union regulates AI ingestion through the risk-based standards of the EU AI Act, as well as through strict lawful bases for processing publicly available training data under the EU GDPR. The EU AI Act distinguishes between four risk categories: unacceptable, high, limited and minimal. These categories define regulatory requirements based on the risks posed by AI systems (Ebers, 2025). In contrast, United States jurisprudence primarily evaluates text and data mining through the fair use doctrine rather than establishing a dedicated copyright exception for artificial intelligence training. The United States relies on the fair use doctrine under Section 107 of the Copyright Act (Du, 2026). When considered alongside Law No. 27 of 2022 on Personal Data Protection (PDP), Law No. 28 of 2014 highlights key statutory gaps in Indonesian law regarding AI data processing and technological adaptation.

Artificial intelligence has substantial potential as an adjunct to protection systems. However, this creates a paradox: during their development, such systems will collect data originating from other sources (Simanjuntak et al., 2024). One type of data gathered, in addition to the limited scientific data available through networks, consists of personal data. Personal data is information that must be protected, particularly sensitive data capable of revealing an individual's identity, which can be classified as non-secular needs or as matters related to respecting oneself. In Indonesia, there are no more detailed provisions beyond Law Number 27 of 2022 on Personal Data Protection; therefore, personal data protection in Indonesia is frequently framed within the domain of human rights protection (Rosadi et al., 2022).

Recent legislative developments in South Korea signal a major shift in data governance relating to artificial intelligence development. Under the amended Personal Information Protection Act (PIPA, Law No. 21445, scheduled for enforcement on September 11, 2026), the use of personal data for training artificial intelligence models is formally permitted, subject to specific conditions. After the bill's passage in August, the statutory framework is expected to be completed by the end of the year. Importantly, the law assigns operational details to subordinate Presidential Decrees, including criteria for evaluating personal data processing conditions, requirements for the adequacy of safeguards, procedures for deliberation and approval, and the extent of simplified compliance. As a result, close monitoring of legislative advance notices is essential, because these enforcement decrees are expected to incorporate practical industry perspectives and broaden commercial data usage.

Within this changing landscape, the regulatory authority of the Personal Information Protection Commission (PIPC) is becoming ever more essential. Although it has historically been criticized for insufficient data breach prevention, the PIPC must now more firmly carry out its enforcement mandate in the context of intensified global debates on AI ethics and high-profile data scraping litigation. Beyond setting lawful limits for AI data use, the PIPC also plays a key part in reducing regulatory arbitrage. In particular, if strict compliance obligations apply only to registered commercial entities and public institutions, unregistered actors could train foundational models and then deploy them via compliant third parties. The broader significance of the forthcoming legal framework is that it may create retroactive accountability for these evasive practices, thereby sealing enforcement gaps across the entire AI development lifecycle.

The form of digital data that is used as training data for Generative AI in large language models is an object with economic value, and it can be protected by copyright, whether it takes the form of images, paintings, sounds, videos, or any other category that copyright protects. In light of this situation, there is a question of copyright ownership regarding digital data used by AI software developers to expand the capabilities of the large language models they own, and whether the digital data is obtained legally or illegally under the relevant copyright provisions.

Personal data can be considered a personal asset because it contains important information that can reveal an individual's track record or even condition, including more strategic data statistics (Spiekermann et al., 2015). Given the lack of data literacy in society, achieving data protection goals for society will be difficult (Park, 2013). A person's digital data can include many forms, such as data stored on their personal devices, personal social media, and even personal data held by the

company where they work. In addition, a person's digital data can be stored by a third party when they grant permission for a company to use their data, for example through digital wallet applications, digital banking services, and ride-hailing services (Acquisti et al., 2016).

Provisions governing the use of digital data obtained from various sources as code input in developing large language models do not explicitly regulate them (Jarota, 2023). In this regard, the ethics of using such data serves as guidance for AI software developers or AI users in minimizing issues that may arise from the use of large language model training data (Hagendorff, 2020). The AI ethics that are currently circulating still mainly focus on AI users in applying AI in society; this situation invites criticism because most AI ethics have not yet focused on AI software developers, particularly concerning the use and acquisition of the training data used (Bender et al., 2021).

There have been several prior studies on AI ethics. One example is research conducted by Nasman, Pudji Astuti, and Dita Perwitasari, titled "Ethics and Accountability of AI Use in Indonesia," which found four ethical aspects that form the basis of AI development: transparency, accountability, fairness, and security and privacy (Nasman et al., 2024). This differs from the present study because we propose human centricity as fundamental ethics. Prior research by Darwance and Marsudi Triatmodjo examined "The Use of AI Based on IPR Law," concluding that AI functions strictly as an object under intellectual property law while also identifying key legal challenges in Indonesia (Darwance & Triatmodjo, 2025). Distinct from their work, this study analyzes the training data used in large language model development under copyright and data protection laws in both Indonesia and South Korea. Martin Jones's study, "Navigating the Privacy Paradox: Balancing AI Advancement and Data Protection in the Digital Age," concluded that there is an intersection point between AI and privacy, so strong privacy protection and enforcement are needed to balance this with the pace of innovation (Jones, 2025). This research also differs from Martin Jones's study because it focuses on how to protect and ensure that the training data provided to generative AI does not violate the minimum requirements for data privacy.

Research by Petar Radanliev concludes that the development of imperative AI poses profound ethical challenges and that protecting the individual requires strong and inclusive knowledge of the individual itself. He identified at least five challenges that must be addressed: privacy, transparency, fairness, and the need to resolve regional challenges and interdependencies (Radanliev, 2025). Radanliev's research differs substantially in that it concentrates on the problems that may emerge, whereas this study proposes an ethical framework that places humans as a fundamental

requirement for addressing those issues. Research conducted by Natalie Zelenka proposes data hazard as a safeguard in data sharing to help build a vocabulary in artificial intelligence, and it emphasizes that people who share their data must be concerned with the data itself in order to minimize the potential for future harm (Zelenka et al., 2025). Meanwhile, in this research, we prioritize an initial screening process and adhere to ethical standards for data based on the principle that all data must first be reviewed by humans.

This research differs from prior studies because we limit the data to the scope of protected data under copyright. In addition, the data question is owned by an individual. This means that a further limitation on the data concerns the data subject, who can identify someone from the perspective of data protection. Analyzing the use of training data for artificial intelligence requires a clear conceptual distinction between legal ethics and positive legal norms. Positive legal norms primarily address formal legal compliance with statutory provisions, which often face a pacing problem linked to rapid technological advancement (Floridi & Cowls, 2019). Under copyright parameters, legal ethics goes beyond the conventional infringement test by assessing respect for the dignity of human creative labor and the principle of fair compensation (Lucchi, 2023). Large-scale extraction of creative works during data ingestion may devalue human authors when it is conducted without moral attribution or equitable remuneration, even if data compression is presented as a non-expensive use (Sag, 2023). Concerning personal data protection, the indicators of legal ethics extend beyond procedural consent by incorporating contextual integrity and informational self-determination. Under this parameter, the ethical indicators require transparency throughout the data pipeline, accountability for the potential of ethical hazards, and mechanisms that ensure data subjects' control over their personal information embedded within parametric artificial intelligence memory (Nissenbaum, 2017).

Based on the previous research mentioned above, each researcher conducted a study on artificial intelligence in each aspect of intellectual property law that is separate from personal data; the research findings take the form of concepts regarding artificial intelligence use and recommendations for national law. Meanwhile, this research complements prior work by focusing on designing an ethical concept for the use of artificial intelligence training data, especially in large language models, across two legal domains: copyright and personal data protection. The ethics of using artificial intelligence training data in large language models, formulated as an initial design grounded in the results of studies of Law No. 28 of 2014 on copyright and Law No. 27 of 2022 on personal data protection, are not sufficient to meet the needs of a comprehensive society. In this digital

era, a proactive approach is required from software engineers and end-users to implement these ethics. This research will develop the ethics of using generative artificial intelligence training data in large language models that can be applied by software engineers and end-users of artificial intelligence.

B. RESEARCH METHODS

This research utilizes normative legal research applying statutory, conceptual, and comparative approaches. Normative legal research is conducted by examining written legal rules and their normative application. The statutory approach analyzes applicable laws and regulations relevant to the research topic, while the conceptual approach provides analytical frameworks to address unregulated legal issues. Furthermore, the comparative approach is employed to systematically contrast domestic legal provisions with foreign and international regulatory frameworks, enabling the identification of statutory gaps, best practices, and avenues for regulatory harmonization regarding Generative AI governance.

The legal materials analyzed consist of primary, secondary, and tertiary materials. Primary legal materials comprise binding legislative instruments, including Indonesia's Law No. 28 of 2014 on Copyright and Law No. 27 of 2022 on Personal Data Protection, alongside key international legal frameworks, namely the European Union Artificial Intelligence Act (EU AI Act), the EU General Data Protection Regulation (EU GDPR), and the United States Copyright Law (Title 17 U.S.C., specifically concerning the Fair Use doctrine in Text and Data Mining). Secondary legal materials consist of academic books, peer-reviewed journal articles, and scientific literature. Tertiary legal materials include legal encyclopedias and dictionaries used to support and clarify concepts as well as analytical outputs obtained from direct prompts administered to Generative AI Large Language Models. The gathered materials were examined through literature studies followed by a qualitative-descriptive legal analysis.

C. RESULTS AND DISCUSSION

Generative AI LLMs require a large amount of data to train their models. This training data is known as a 'big data corpus', and the deep learning architecture connects this data in order to predict the next output. Generative AI LLMs use all kinds of data available on the network, including personal data that is subject to legislation and creative works that already exist and are protected by copyright. In these cases, using creative work by Generative AI LLMs causes substantial issues with the exclusive rights to the moral and economic rights of creative work.

From the perspective of personal data protection, all personal data is subject to protection under personal data protection law, including both specific and general personal data. Personal data is

closely related to privacy, which is a human right. The protection of personal data includes how the data is obtained, processed and used. This condition can be bypassed by the industry-related Binding Corporate Rules (BCRs), which eliminate the need for data gathering, processing, transfer and use, and are standardised in relation to national law, even when data is transferred across borders.

On the other hand, the need for human expertise is inevitable. The gathering, processing, development and delivery of results of Generative AI LLM need to be overseen by humans. Data privacy oversight also places humans at the top of the controller hierarchy, mitigating any risk to companies and individuals. This is what we propose as the human-centric principle in AI data training.

1. Generative AI LLM Training Data from a Copyright and Personal Data Perspective

In this research, we divide the discussion into two main discussions, namely data training from copyright perspective and data protection perspective. For the second discussion, we do examine and elaborate more in the field of AI ethics with the complementary data from Generative AI LLM itself.

a. Data Training from Copyright Perspective

AI is the development of algorithms in computer programs. Initially, computer programs were created with algorithms to run programming languages as input instructions, the output of which was determined by the algorithm in the program. However, computer program technology now uses more complex algorithms because the algorithms are written as more complex input instructions, and the output is more varied. A study by Heliza posits that AI training utilising voice data consistently carries the potential to derive from voices (Heliza et al., 2025). However, a voice itself is generally not a copyrighted work; it falls under related or personality rights.

Ethical considerations must comply with prevailing legal regulations (Wahyuni & Turisno, 2019). When utilising copyright-protected data, various factors must be considered, such as ethics, integrity and individual circumstances, all of which influence how such data is processed and used. Digitalisation increases vulnerability to the unauthorised appropriation and misrepresentation of protected intellectual property. Moreover, given that Indonesia has enacted copyright legislation, this statutory framework requires further development in the digital era to establish appropriate benefit-sharing mechanisms and proper copyright protection (Roisah et al., 2025). This means a code of ethics serves as a preventive measure and a vital component of corporate governance.

The ability applied to AI is machine learning, which uses datasets as training data to increase AI performance. A large amount of relevant training data, also known as big data, is called a corpus. Machine learning focuses on developing algorithms

and models that enable computers to learn from data, providing them with the ability to learn without being explicitly programmed. In this context, machine learning can be understood as a system's adaptive ability to improve performance based on growing data (Bintoro et al., 2024).

Machine learning has a specific derivative ability known as deep learning. Deep learning uses artificial neural networks to process more complex data and can automatically scan and extract information from large, varied and unstructured training data (Sarker, 2021). Deep learning is key to processing large-scale datasets or big data applied to an LLM. In this context, big data or a corpus that becomes AI training data with machine learning and deep learning capabilities is the most important factor in improving Generative AI performance, particularly for LLMs (Zhao et al., 2026). This leads to more advanced Generative AI LLMs because, naturally, the more big data that is used, the more it will be in line with the scaling laws of large language models (Kaplan et al., 2020).

Big data or corpuses used as training data are digital files in the form of text, images, videos or multimedia. These digital files intersect most dominantly with copyright, one of the IPR classifications. If training data in the form of digital files is a work resulting from human thought, it is automatically protected by copyright law when it meets the requirements for copyright protection, i.e. fixation, form, and originality.

In Indonesia, copyright law protects the exclusive rights of creators over the results of their intellectual work. This gives creators moral and economic rights over their creations, allowing them to control how their work is used by others. The subject matter of copyright in the fields of science, art and literature is explicitly listed in Article 40, Paragraph 1. Generative AI in the form of LLMs is an object protected by copyright as a computer program. Conversely, datasets or corpora used to train LLMs are potentially works protected by copyright, as listed in Article 40 of Law No. 28 of 2014, even if they are in digital form. The copyright issue with generative AI LLM training data lies in how the data is utilised as code input to create output. These data may have been originally protected by copyright, but were stored and used for developing LLM performance without permission, effectively changing the form of the copyrighted work to create new works, as Generative AI does.

The problem caused by using creative works in the form of digital files utilized as training data for generative AI especially lies in the aspect of reproduction of creative work. Essentially, the use and even reproduction of a creative work must obtain permission from the creator or the copyright holder of the work, and by IPR law, there is a prohibition against acts of reproduction and/or commercial use of creations. Reproduction in copyright defined as a process, action or way of

duplicating one copy of a creation or more in any way and in any form permanently or temporarily. The term of commercial use is defined as utilization of a work of creation for the purpose of obtaining economic profit from various sources or paid work (Gema, 2022). The use of creations as training data for Gen AI or LLM can be categorized as an act of reproduction of creative work. Properly, AI software engineers should ask for permission from the creator/copyright holder of creative work to be copied and duplicated for the purpose of developing AI or LLM technology as training data.

The autonomous Gen AI does not satisfy the al 'aql requirement, meaning the human safeguards Gen AI (Saparwadi et al., 2026). Meanwhile, even though data is already digitalised, there is a need to strengthen human resources and human literacy (Silviana et al., 2025). Nevertheless, both natural persons and legal entities must adhere to applicable codes of conduct for compliance, risk mitigation and enhanced transparency, with the aim of cultivating responsible practices (Turisno et al., 2024).

In the context of the standard of copyright for works created using generative AI and LLM output, the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission maintain the fundamental stance that copyright cannot be granted to standalone generative AI and LLM outputs. However, they are increasingly clarifying the requirements for what constitutes a recognisable human creative contribution. The core judgement criteria are based on whether the creator intended the desired outcome and directed the overall process, and on whether the output could be precisely guided in accordance with the creator's intentions through the inputs. Consequently, simply typing text prompts multiple times is generally considered to demonstrate low control and predictability due to high randomness, making it difficult to secure copyright based on prompts alone. However, if AI-generated images undergo inpainting (modifying specific areas), multi-step refinement, precise combination with existing human-made works or direct post-processing tasks such as adjusting colour, composition and arrangement, copyright registration is permitted strictly for the portions contributed by human labour.

From the perspective of copyright limitations such as fair use in the US and reasonable interest in Indonesian copyright law, the use of creative works as training data is difficult to classify as utilisation that meets the limitations of copyright. This is because generative AI, especially LLMs, is software that is not purely developed for educational and research purposes. However, there are several relevant cases in the US relating to this issue: Thomson Reuters v. Ross (2025, no fair use); Bartz v. Anthropic and Kadrey v. Meta (2025, fair use); and GEMA v. OpenAI (Munich, November 2025, infringement through memorised lyrics). Generative AI and LLMs charge a subscription fee

to users to access their software. Although some Generative AI and LLMs provide free services, the data provided by users through command prompts and uploaded data used to provide instructions for Generative AI or LLMs are used as input data. This occurs without the user's consent and is used to develop Generative AI and LLM software.

Using creative works as training data for generative AI, especially LLMs, raises issues of copyright, particularly with regard to the exclusive rights to the moral and economic aspects of creative works. In terms of moral rights, there is a lack of transparency surrounding the data used for training, and sometimes the generated output dominantly resembles the creative work of others. This suggests a lack of attribution or respect for the moral rights of others' creative works. In terms of economic rights, the majority of Generative AI, especially LLMs, generate economic value through subscription fees from users who want to use generative AI or LLM technology, but these economic benefits are not shared with the creative works that have been used as training data. This leads to Generative AI having the highest potential to infringe copyright by reproducing creative works.

The use of training data, particularly through text and data mining (TDM), in the development of generative AI and LLMs requires careful analysis due to structural disparities across jurisdictions. In the EU, TDM exceptions are codified under Articles 3 and 4 of the DSM Directive (2019/790), which incorporate rightsholder mechanisms alongside the obligations of the AI Act for general-purpose AI. These obligations cover copyright policy and training data summaries under Article 53 and will be effective from 2 August 2025. The European Union uses the three-step test derived from Article 9(2) of the Berne Convention to regulate general copyright exceptions.

Zhenyi Gao provides a comparative legal analysis of the Text and Data Mining standard system across four elements of fair use under article 17 of US Copyright Law. In terms of judgment standard, the U.S. applies the four elements. Furthermore, the U.S. regime operates under a loose judgment scale and grants larger discretion to judges, leading to an open scope of exemption (Gao, 2024).

Fair use is a legal concept that limits a creator's exclusive rights, allowing reasonable and transformative uses such as education, research, criticism, news reporting, and parody to balance copyright protection with the public interest (Yulianti & Roisah, 2026). In brief, the fair use doctrine permits restricted use of copyrighted works for education and research without requiring the copyright holder's consent (Jened et al., 2026). Addressing transformative use of US Fair Use, statutory compensation frameworks for gen AI training should align with commercial use and output transformativeness. Highly transformative or researchoriented models generally do not require

levies, as they present little risk to original authors. However, commercial general purpose AI Generative with LLM model with low transformativeness need financial levies and structural caps to prevent market dilution and protect copyright holders (Lucchi et al., 2026). Applying transformative US fair use to AI data relies on a non-cumulative balancing of four elements factors: the purpose and character of the use, the nature of copyrighted work, the amount used, and the effect on the potential market. Data training of AI for LLM primarily asserts a transformative purpose by converting expressive content into non-expressive statistical data for pattern extraction. Because these factors are non-cumulative meaning no single factor is dispositive and they do not all need to be satisfied simultaneously a strong showing of transformativeness and minimal market displacement can legally outweigh the complete ingestion of creative works.

The intersection of AI data training and IPR remains a subject of intense regulatory and industry debate in South Korea, with recent legislative and policy trends increasingly accommodating the operational needs of the AI industry. Legal professionals and content stakeholders remain sharply divided over the legal boundaries of utilizing proprietary copyrighted material to train generative models. Under emerging government guidelines, the judicial scope of the "fair use" doctrine has significantly narrowed. Specifically, ingesting full text news articles, visual art, or literary works without authorization to power competing summarization or content generation platforms is deemed to directly impair the economic viability of original creators, thereby exceeding the permissible thresholds of statutory fair use. Consequently, to reconcile these market tensions, media outlets, webtoon and web novel publishers, and music copyright management organizations are actively spearheading negotiations to establish structured licensing frameworks and equitable remuneration models that mandate fair compensation for rights holders whose assets serve as foundational training data.

In creative fields and the technology market, complementary measures are strengthened to transparently disclose AI usage and verify rights. Enhanced Requirements for Copyright Registration, when registering AI assisted works with the Korea Copyright Commission, applicants must submit more than just the final outcome they are required to provide materials demonstrating creative contributions, such as the scope of AI utilization, human modification and processing workflows, and prompt revision history. Adoption of Watermarks and Generation Labels. Discussions continue regarding the implementation of standards like C2PA (Coalition for Content Provenance and Authenticity) or legal obligations to label AI-generated content to prevent issues like fake news and unauthorized appropriation.

b. Personal Data in Generative AI LLM

The concept of personal data is equated with data that can identify a person, including data scattered across the network and data stored on their own device. Personal data is closely related to who can access it, and generally only the owner can access their own data (Erdos, 2022). Personal data has evolved alongside technological developments, particularly in the 5.0 technology era where the internet has created a free and unlimited market (Wibowo et al., 2024). Consequently, integrity is a paramount ethical principle that must be prioritised (Samiri et al., 2025).

One study shows how data can shape business, government and social interactions. However, the large amount of data collected, processed and used could potentially cause loss to individuals because it is often not taken directly from them, but from things they are not aware of. This is the primary reason why data is regulated by the government (Abrams, 2014). According to the Warren & Brandeis theory of 'the right to be let alone', which was later developed by Alan Westin into the concept of privacy as control over when, how and to what extent information is shared, data is either allowed or not allowed to be known by others based on the will and concern of the data owner (Marky, 2023). This concept is what we know today as privacy.

Conceptually, everyone as human-being is protected by a right known as the right to privacy. The right to privacy protects every individual who lives 'the living things' without exception, even if that person is a politician or celebrity. In some countries, the right to privacy even protects those who have died with the goal to protect their family privacy. In the concept of data transfer or transmission, from one country to another country, a similarity in the level of personal data protection regulation is needed. However, data transfer within one group of companies is permitted even if the two companies are domiciled in two different countries. The permission is based on the Binding Corporate Rules ("BCR") which consider the group as one legal entity (Bakar, 2025).

Indonesia has regulated personal data protection through Law Number 27 of 2022 concerning Personal Data Protection ("PDP Act") which take effect on October 2022. In the Academic Draft of the UU PDP, it is explained that one of the reasons for the birth of the UU PDP is a form of recognition of privacy as Human Right. The concept of privacy is equated with 'trespass', namely entering other people's residence without permission. In a concept like this, the consent of a data owner is needed so that the data can be accessed by others (Datapribadi.id, n.d). This concept are very similar to Islamic concept mentioned in QS An-Nur 27 – 28, which said "do not enter the house other than you own until you have asked permission and greeted its occupant"

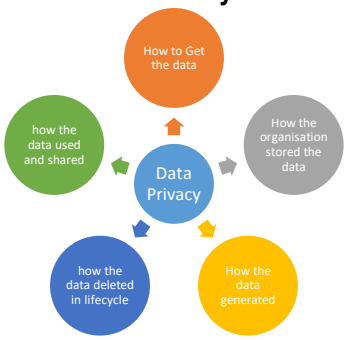
and “if you find no one at that home, do not enter it until you have given permission”.

The PDP Act divides personal data into two main categories: specific personal data and general personal data. General personal data is divided into several types, including full name, gender, nationality, religion, marital status and data that can be used to identify an individual. On the other hand, specific personal data includes data that can describe physiological and physical characteristics, as well as unique identification data such as biometric data, and also covers health, genetic, criminal record, children's and personal financial data (JDIH BPK Database Regulations, 2022). The processing of personal data, both specific and general, requires valid and full consent for access and processing.

In terms of processing personal data, it is normal for a corporation to access and process personal data outside of its original countries if it has business there. The concept of BCRs emerged in Europe in response to the need to regulate multinational companies operating across borders. This concept indirectly eliminates the obligation to report and supervise data transfers between countries on the grounds of similar regulatory standards. However, this principle creates loopholes, such as uncontrolled data transfer, and transferred data can be audited by data auditing institutions in the recipient country. The BCR concept also creates an obligation for a competent data protection authority to exist at company, country and data recipient country levels. Thus, BCRs do not 'eliminate the obligation for reporting and supervision'; they require data authority approval and remain auditable. A BCR does not make the group 'one legal entity' or predate the GDPR, which codified BCRs in Art. 47 (Rees & Hodgkinson, 2007).

In the era of AI development, BCRs play an important role in the data transfer process, particularly with regard to ensuring the safe transfer and management of data, and establishing compliance obligations in the realm of data protection (Compagnucci et al., 2024). BCRs can be used as an internal monitoring mechanism to complement the independent data protection supervisor structure mandated by the UU PDP. The aim is to ensure that data protection standards are suitable and applicable at country and company levels. Additionally, it can be used to improve data protection by reviewing processes and/or data inputs (Tarikh & Bahri, 2025; Hakim et al., 2023). Based on prior research and the UU PDP, researchers have chosen to highlight the importance of data safety and privacy.

Figure 1. Illustration of Data Protection for Privacy and Security



Related to privacy protection in AI, ASEAN region already made a guide on AI Governance and Ethics which has been published in 2024. In the ASEAN Guideline, there are several things that need to be considered in AI Construction. Starting from designing, developing, and implementing AI. The construction of AI Ethic embraces how the data is collected, stored, processed, and deleted from AI storage is required to adapt existing data regulations, data protection compliance, and existing ethical principles. From the ASEAN Guide on AI Governance and Ethics, every organization including both government and private sector are recommended to perform periodic checks, get explicit and valid consent from someone whose personal data will be taken, stored, processed, and forwarded (ASEAN Secretariat, 2024). Researchers illustrate the correlation between benefits and risk in data protection as follows:

Figure 2. Illustration of a Personal Data Consent Generated



A recent study found that privacy has six elements that influence a person's decision to consent to sharing or allowing access to their personal data. These elements are: human agency; the use and abuse of privacy; AI transparency and opacity; the weaponisation of information; cyberbullying and privacy issues; and the enforcement of privacy strategies. If the benefits are perceived to outweigh the risks, then personal data or privacy is more likely to be shared. Conversely, if the risk is perceived to be greater, personal data will be increasingly protected, or at least not shared or accessible (Akakpo et al., 2025).

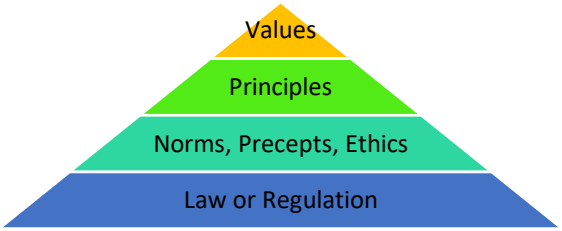
A key consideration is the necessity of harmonising cross-sectoral regulations to designate a supervisory authority responsible for overseeing data governance. Within the data protection framework, both the state and data-aggregation corporations play pivotal roles. While data protection governance has been established through comprehensive national legislation in several jurisdictions, in Indonesia it tends to be sectorally fragmented. Robust data protection requires not only statutory regulations, but also continuous capacity building for individuals involved in data processing (Kumalaratri & Yunanto, 2021).

2. Ethics of Generative Artificial Intelligence Large Language Models Training Data Usage

In this increasingly democratic and fast-paced life, ethics has its own place. Although ethics are not always made into legal norms, they are often considered a pillar of modern life. Some people see ethics as a role model, giving basic guidance in the same way as leadership or values. In a healthy democratic world, there will always be agreement on ethics that contain values of honesty, fairness and responsibility. However, ethics is often considered a grey area between lived values and values that have been turned into norms (Sapan & Santiago, 2025).

Some research suggests that ethics is often associated with awareness and compliance. It is considered a fundamental basis for maintaining trust, integrity and fairness in social life, even though it is often considered to be in paradox with legal rules because it does not provide clear guidance (Kodner et al., 2007). Other research suggests that ethics are important in all activities, particularly business activities, where they are often considered less applicable as a model for providing sanctions (Dutta et al., 2022). This may be because ethics and law operate at different levels: ethics is where law is found, while law is the formalised version of ethics with sanctions attached. A sociological approach to ethical violations will yield a degree of reproach appropriate to the conduct, or even a level of condemnation beyond the act committed. At the level of law, however, violations lead to sanctions (Heryansyah, 2018).

Figure 3. The Hierarchy of Values, Principle, Ethics, and Law



Source: (Cohen, 1981; Magnis-Suseno, 1993; Murtadho et al., 2024)

To bridge these distinct operational levels systematically, recent literature establishes a structured continuum that organizes normative governance into an empirical hierarchy spanning values, principles, ethics, and law. These overarching principles are then translated into domain-specific ethical guidelines and protocols, such as responsible data ingestion practices, ethical consent mechanisms, and contextual data protection standards. When ethical breaches pose systemic risks requiring binding enforceability and formal sanctions beyond social reproach, these ethical solidify into positive laws and statutory rules that based on broadly acceptable principles to guide the people and processes responsible development (Mittelstadt, 2019).

In this research, researcher's places discussion by gathering ethics considered appropriate to formulate the collection, management, and delivery of answers based on data on Generative AI LLM. Ethics in computer

technology is in the form of the nature and social impact of computer technology and the formulation and justification of policies appropriate for the ethical use of technology.

Ethics of an AI is faced with three things, namely toward the AI itself, Humans, and Society. AI Ethics toward AI is in the form of principles of AI development to interact with other AI. AI Ethics with humans in the form of principles of AI development to interact with humans ethically. AI Ethics toward society on principles of AI development as an ethical function in society construction (Siau & Wang, 2020).

AI ethics is a set of values, principles, and techniques that apply widely accepted standards of what is 'wrong' and what is 'right' to guide moral behavior in the development and use of AI technology. Values, principles, and techniques intended to motivate morally acceptable practices and set tasks and obligations to produce an ethical, fair, and safe AI technology. There are several potentials that can be caused by the massive use of AI technology systems such as bias and discrimination, denial of individual rights, non-transparent results, invasion of privacy, isolation and social disintegration, and unsafe results (Leslie, 2019).

One elaboration of AI ethics was defined by an American multinational technology company named International Business Machines (IBM) which focuses on hybrid cloud development and artificial intelligence. According to IBM, ethics is a set of moral principles that help distinguish between right and wrong, and AI Ethics is a multidisciplinary field that studies how to optimize the beneficial impacts of AI while reducing risks and adverse outcomes. IBM sets AI ethical principles that serve as a guide for experimentation and algorithm design on AI, namely the principle of respecting humans, the principle of benefit, and the principle of fairness (Clark & Kosinski, 2026). As comparative guidance, AI ethics guidelines have been made comprehensively:

Table 1. Comparison of Major AI Ethics Norms

Key item	Republic of Korea: Korea AI Ethics Principles (2020)	EU High-Level Expert Group (AI HLEG): Ethics Guidelines for Trustworthy AI (2019)	Japan: Social Principles of Human-Centric AI (2019)	OECD: Recommendation of the Council on Artificial Intelligence (2019, amended 2024)	UNESCO: Recommendation on the Ethics of Artificial Intelligence (2021)	ASEAN: ASEAN Guide on AI Governance and Ethics (2024)	G7*: Hiroshima Process International Guiding Principles (2023)
Human dignity and human rights	● [Value] Human dignity	●	●	●	●	△	△
Social public good and	● [Value]	●	●	●	●	△	△

Key item	Republic of Korea: Korea AI Ethics Principles (2020)	EU High-Level Expert Group (AI HLEG): Ethics Guidelines for Trustworthy AI (2019)	Japan: Social Principles of Human-Centric AI (2019)	OECD: Recommendation of the Council on Artificial Intelligence (2019, amended 2024)	UNESCO: Recommendation on the Ethics of Artificial Intelligence (2021)	ASEAN: ASEAN Guide on AI Governance and Ethics (2024)	G7*: Hiroshima Process International Guiding Principles (2023)
well-being	Public good of society						
Sustainability	● [Value] Sustainability of humanity	●	●	●	●	△	●
Human-centeredness and human agency	● [Principle] Human centeredness	●	●	●	●	●	△
Privacy and data protection	● [Principle] Privacy protection	●	●	●	●	●	△
Fairness, non-discrimination and inclusiveness	● [Principle] Fairness and inclusiveness	●	●	●	●	●	△
Accountability	● [Principle] Accountability	●	●	●	●	●	△
Safety and security	● [Principle] Safety	●	●	●	●	●	●
Reliability and robustness	● [Principle] Reliability	●	△	●	△	●	△
Transparency and explainability	● [Principle] Transparency	●	●	●	●	●	●

*The table shows which key ethics items each of seven frameworks covers:
● marks an item that is addressed,
△ one that is only partly addressed.
** (G7) The Hiroshima Process document is closer to a code of conduct covering the specific actions and measures needed to manage risks from advanced AI, so the range of concepts that directly correspond to the other ethics norms is relatively limited.

Source: (AI Ethics Communication Channel, 2020); (KLRI, 2025) (European Comission, 2019); (Regulations.ai, 2019); (OECD Legal Instruments, 2024); (UNESCO, 2021); (ASEAN Secretariat, 2024); (G7 G20 Document Database, 2023)

In 2021, the European Union published an ethical design for available AI. Explicitly, in that design a trustworthy and usable AI must meet six basic ethics arranged in levels. These six levels of basic ethics must be applied and followed by a minimum base indicator test. From the European Union hierarchy design, these basic ethics are constructed as follows:

Figure 4. Illustration of AI Ethics Hierarchy



Source : (European Comission, 2021)

When testing whether the six levels of basic AI ethics were applied in Generative AI LLMs, the researchers also tested an input prompt in the form of a questionnaire. These input prompts consisted of a basic question: 'How were the six levels of AI Ethics Principles applied in [name of Generative AI LLM]?'. The researchers selected several Generative AI LLMs as sources of AI ethics application data: ChatGPT, Gemini AI and Meta AI. These Generative AI LLMs were selected based on popularity and accessibility between 1 January and 31 October 2025. Popularity and accessibility data were generated using ChatGPT prompts. ChatGPT data was used because the other Generative AI LLMs do not have any data resources. The prompt used to process the data employed a simple formula as follows:

$$\text{Popularity (\%)} = \frac{\text{Number of Platform Users}}{\text{Total Users of All AI Platforms}} \times 100$$

We chose six Generative AI LLM that is already well known, namely ChatGPT, Gemini AI, Meta AI, Claude AI, Perplexity, and Grok AI. Researchers use data generated from ChatGPT, as disclaimer, we must note that this data might differ from other resource and or researcher, this data may change depending on the prompt, the time of input prompt. This data was chosen because during the research, no institution has been found that can provide these user figures other than Generative AI LLM itself. We use prompt in ChatGPT as follows:

"Data was obtained from ChatGPT with input prompt: Create a table containing statistical data on users of ChatGPT, Gemini AI, and Claude AI. The data must come from a reliable source or accessible OpenAI data including internal data from OpenAI, ChatGPT, Gemini AI, and Claude AI. The popularity rate shall be calculated by dividing the total active users of the specific platform by the aggregate number of users across all identified AI Platforms, multiplied by one hundred (100). The columns are as follows: month | number of users | percentage | market share. You can create three separate data tables or improvise only for the tables."

In general, from data generated by ChatGPT, researchers chose three most popular and accessed as follows:

Table 2. Generative AI LLM Access and Popularity Data

AI Model	User access in January 2025	User access in October 2025	Total Growth (Jan-Oct)	Estimated Market Share
ChatGPT	~350-400 million	~800 million	~100-128%	≈60-64%
Gemini AI	~430 million	~650 million	~51%	≈13-15%
Meta AI	~1,0 billion	~1,1 billions	~24%	≈18-20%

Source: (Search Engine Land, 2026; Tangf, 2025; Vaswani, 2025)

From many forms of ethics guidelines, researchers chose four ethics principles in Generative AI LLM. Four principles tested were fairness, security and safety, privacy and data protection, and accountability and transparency. From the three Generative AI LLM chosen, these are the results:

Table 3. Generative AI LLM Ethical Principles Implementation Test

Ethical Principle	Implementation on Gemini AI	Implementation on ChatGPT	Implementation on Meta AI
1. Fairness and Bias Handling (Fairness)	Gemini running a strict test to detects or reduces algorithmic bias to deliver a response. This is a delivery from using diverse training data and filters for stereotypes or discrimination, mainly in individual, ethnic, and gender representation.	Avoids giving a bias answer related to gender and ethnicity. Avoids using a multiple interpretations words, and adjusts to user location (country) terms.	Meta AI was constructed to avoids bias and discrimination, also ensuring equal access for all kind of users.
2. Safety and Security	Gemini made a guardrails against some content generated, namely possible	Rejecting all request prohibited by law and unethical request.	Creating a design to be safe without risk to users.

Ethical Principle	Implementation on Gemini AI	Implementation on ChatGPT	Implementation on Meta AI
	dangerous and wrong facts. This guardrails including illegal advice, hate speech, explicit content, exploitation content, dangerous advice, and hallucination. For safety, Gemini AI increasing data to avoids incorrect or factually misleading advice.		
3. Privacy and Data Protection	Implement strict user data privacy. Data generated from google workspace will not included as training data unless explicitly permitted. The permission can be changed by user in setting menu or explicit prompt. Gemini AI using a standardized safety and control as embedded to other Google Workspace	ChatGPT will not use personal data without explicit permission. Fo registered user, the data generated from user question will be used as data personalization and for fine-tuning by default. In some prompt we tested, ChatGPT refuse to deliver a picture of politically exposed person or famous person.	Meta AI ensure personal data will not use as training data and opt-out for training data. Meta AI giving a optional for user whether the conversation will be used as training data or not with conclusion they will anonymizes data.
4. Accountability and Transparency	Gemini explains how they work and process the data and will disclose that it is AI generated not human generated. Users can also give feedback on inaccurate responses delivered from Gemini AI. Gemini provide source citations for facts. In	ChatGPT explains the source and limitations of datasets. They will deliver the answer from input prompt with some process: *Encoding (embedding)-> Prediction -> Decoding -> Alignment Filtering*. Chat GPT does linguistic pattern	Meta AI will make extra effort to disclose and give transparent system decisions and mechanisms. They have mechanism for accounts for errors and are responsible for errors they made.

Ethical Principle	Implementation on Gemini AI	Implementation on ChatGPT	Implementation on Meta AI
	this response, the fact should be seen as the answer of a prompt.	analysis before delivering answers. They intend to give educative advice and sometimes they will not include the source of answer.	

Source: Generative AI LLM Test Subject: (OpenAI, 2025), (Google Gemini, 2025), (Meta Platforms, 2025)

From the table above, we know that Generative AI LLM has its own principles. The regulation can be a "game changer" in industry. Several points might be considered to ensure the safety of users and potential social behavior, such as inserting AI principles into respective laws, enacting supervisory board, establishing personal data protection officer, and establishing government specific supervisory for Generative AI LLM (Armiwulan et al., 2024). AI emerged as powerful tool with the potential to enhance decision-making in each domain even the challenges and limitations are still exist (Hariyanti et al., 2023).

D. CONCLUSIONS

This research shows that generative AI LLMs convert all input data, prompts and corrections into a means of developing the model further. This requires ethical AI supervision from the outset, covering everything from data gathering and processing to delivering answers, in order to overcome security and privacy gaps under the Personal Data Protection Law and lawful bases such as explicit consent, contract, legal obligation, vital interest, public interest and legitimate interest. This regulatory boundary directly impacts the permission or prohibition of data crawling from search engines and mandates human supervision, even within corporations utilising binding corporate rules. Furthermore, in terms of intellectual property rights (IPR) under the Copyright Law and the US Fair Use doctrine, copyright recognition for AI outputs is strictly limited to human creative contributions based on control and predictability, meaning that standalone text prompts lack sufficient human authorship. However, substantial manual edits, inpainting or structured compilations qualify. As claims of transformative training use under fair use are heavily constrained by economic harm to original creators in the market, governance must shift towards explicit licensing, fair compensation and technical verification standards, such as

provenance labelling. This can be achieved by establishing a system that assists humans in supervising processing flows, balancing responsibilities between developers and end users, and aligning AI innovation with legal accountability and privacy rights.

As the research progresses, we conclude that some basic ethics must be proposed for stakeholders, particularly in Indonesia. The proposed ethical framework for generative AI large language model (LLM) stakeholders outlines a twofold model of shared responsibility, divided between software engineers and end users. On the development side, software engineers must establish human supervision by appointing data protection officers or expert teams. They must also maintain transparency in the user interface regarding data permissions, provide clear attribution for training data sources, specify licensing terms for works generated by LLMs, implement safeguards to limit outputs that could harm legal subjects, and ensure that terms and conditions are accessible in layman's terms and in formats that are accessible to people with disabilities. They must also operate a responsive complaints mechanism that reports investigation results back to users. Conversely, end users are ethically obligated to select platforms with explicit intellectual property and privacy terms; comply strictly with output licences; refrain from using AI tools for unlawful or reprehensible activities; actively report suspected misuse; and participate in ongoing oversight by providing feedback to address potential violations of privacy, intellectual property rights and algorithmic bias. Together, these complementary duties form an integrated governance system that balances technical accountability with responsible user stewardship.

Based on our existing research, we know that Generative AI LLMs use all data, datasets, input prompts, conversations and corrections from users to develop the Generative AI LLM further. Establishing human supervision is also essential to ensure compliance with copyright and privacy laws and to overcome security and privacy gaps. Explicit user consent and freedom of choice remain the primary boundary for every Generative AI LLM.

REFERENCES

JOURNALS

Abrams, M. (2014). The Origins of Personal Data and its Implications for Governance (SSRN Scholarly Paper No. 2510927). Social Science Research Network. https://doi.org/10.2139/ssrn.2510927

Acquisti, A., Taylor, C., & Wagman, L. (2016). The Economics of Privacy. *Journal of Economic Literature*,*54*(2),442–492. https://doi.org/10.1257/jel.54.2.442

Akakpo, A., Gyasi, E. A., Oduro, B., & Akpabot, S. (2025). Artificial Intelligence (AI) and the Future of Information Privacy: Expert

Viewpoints. *Journal of Global Information Management*,*33*(1),1–25. https://doi.org/10.4018/JGIM.383050

Andini, K., Permata, R. R., & Ayu, M. R. (2021). Pelindungan Hak Cipta Pada Penggunaan Gambar Di Internet Yang Dijadikan Desain Pada Fitur Filter Cerita (Stories) Aplikasi Instagram Secara Tanpa Hak Untuk Tujuan Komersil. *Widya Yuridika: Jurnal Hukum*, *4*(2). https://doi.org/10.31328/wy.v4i2.2804

Armiwulan, H., Rahman, R. A., Prabowo, V. N., & HajdÃº, J. (2024). Artificial Intelligence and Its Challenges To Elections In Indonesia: A Legal Analysis. *Jambura Law Review*, *6*(2), 264–285. https://doi.org/10.33756/jlr.v6i2.24243

Cohen, H. (1981). Kelsen's Pure Theory of Law. *The Catholic Lawyer, 26*(2). https://scholarship.law.stjohns.edu/tcl/vol26/iss2/4/

Darwance, D., & Triatmodjo, M. (2025). Penggunaan Artificial Intelligence Berdasarkan Hukum Hak Kekayaan Intelektual: Pengaturan di Indonesia dan Tantangannya Secara Global. *Jurnal Magister Hukum Udayana (Udayana Master Law Journal)*, *14*(3), 665–693. https://doi.org/10.24843/JMHU.2025.v14.i03.p07

Dutta, D., Mishra, S. K., & Budhwar, P. (2022). Ethics in competency models: A framework towards developing ethical behaviour in organisations. *IIMB Management Review*, *34*(3),208–227. https://doi.org/10.1016/j.iimb.2022.10.002

Ebers, M. (2025). Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU's AI Act. *European Journal of Risk Regulation*,*16*(2),684–703. https://doi.org/10.1017/err.2024.78

Erdos, D. (2022). Identification in personal data: Authenticating the meaning and reach of another broad concept in EU data protection law. *Computer Law & Security Review*, *46*. https://doi.org/10.1016/j.clsr.2022.105721

Floridi, L., & Cowls, J. (2019). A Unified Framework of Five Principles for AI in Society. *Harvard Data Science Review*, *1*(1). https://doi.org/10.1162/99608f92.8cd550d1

Gao, Z. (2024). Judicial Judgment Standard of Generative Artificial Intelligence and Training Data from the Perspective of Fair Use. *Transactions on Social Science, Education and Humanities Research*, *7*(1), 17–30. https://doi.org/10.62051/2wk20c78

Gema, A. J. (2022). Masalah Penggunaan Ciptaan Sebagai Data Masukan Dalam Pengembangan Artificial Intelligence Di Indonesia. *Technology and Economics Law Journal*,*1*(1). https://doi.org/10.21143/TELJ.vol1.no1.1000

Guo, J., Wang, M., Yin, H., Song, B., Chi, Y., Yu, F. R., & Yuen, C. (2025). Large Language Models and Artificial Intelligence Generated Content Technologies Meet Communication Networks. *IEEE Internet of Things Journal*, *12*(2),1529–1553. https://doi.org/10.1109/JIOT.2024.3496491

Hagendorff, T. (2020). The Ethics of AI Ethics: An Evaluation of Guidelines. *Minds and Machines*,*30*(1),99–120. https://doi.org/10.1007/s11023-020-09517-8

Hakim, G., Jabalnur, Haris, O. K., Ruliah, Sukring, & Mohammad, M. (2023). Analisis Perbandingan Hukum Mengenai Regulasi Perlindungan Data Pribadi Antara Uni Eropa dan Indonesia. *Halu Oleo Legal Research*, *5*(2),443–453. https://doi.org/10.33772/holresch.v5i2.338

Hariyanti, E., Janeswari, M. B., Moningka, M. M., Aziz, F. M., Putri, A. R., Hapsari, O. S., Sutha, N. A. A. D., Sinaga, Y. A. A., & Bendesa, M. P. (2023). Implementations of Artificial Intelligence in Various Domains of IT Governance: A Systematic Literature Review. *Journal of Information Systems Engineering and Business Intelligence*, *9*(2), 305–319. https://doi.org/10.20473/jisebi.9.2.305-319

Heliza, C. C., Santoso, B., & Sutrisno, P. A. (2025). Problematika Penerapan Doktrin Fair Use Dalam Penggunaan Artificial Intelligence (Ai) Untuk Duplikasi Suara Penyanyi: Studi Kasus Legalitas Lagu “Heart On My Sleeve” oleh AI. *Diponegoro Law Journal*, *14*(3). https://ejournal3.undip.ac.id/index.php/dlr/article/view/51660

Jarota, M. (2023). Artificial intelligence in the work process. A reflection on the proposed European Union regulations on artificial intelligence from an occupational health and safety perspective. *Computer Law & Security Review*,*49*.https://doi.org/10.1016/j.clsr.2023.105825

Jened, R., Raharjo, I. S., Kumar, A., Rathi, A., Sari, B. N. I., & Zulfah, S. (2026). Copyright Protection of Computer Software: A Comparative Study of India and Indonesia. *Law Reform*, *22*(1), 113–149. https://doi.org/10.14710/lr.v22i1.64235

Jones, M. (2025). Navigating the privacy paradox in a digital age: Balancing innovation, data collection and ethical responsibility. *Journal of Ethics in Entrepreneurship and Technology*,*5*(1),2-13. https://doi.org/10.1108/JEET-12-2024-0040

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020). *Scaling Laws for Neural Language Models* (arXiv:2001.08361).arXiv. https://doi.org/10.48550/arXiv.2001.08361

Kumalaratri, G., & Yunanto, Y. (2021). Urgency Of The Personal Data Protection Bill On Privacy Rights On Indonesia. *Jurnal Hukum Unissula*,*37*(1),1–13. https://doi.org/10.26532/jh.v37i1.13604

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Zenodo. https://doi.org/10.5281/zenodo.3240529

Liang, H., Zhao, Z., Han, Z., Qiang, M., Ma, X., Zeng, B., Cai, Q., Li, Z., Tang, L., E, W., & Zhang, W. (2026). *Towards Next-Generation LLM Training: From the Data-Centric Perspective* (arXiv:2603.14712). arXiv. https://doi.org/10.48550/arXiv.2603.14712

Liu, Y., Cao, J., Liu, C., Ding, K., & Jin, L. (2025). Datasets for large language models: A comprehensive survey. *Artificial Intelligence Review*,*58*(12),403.https://doi.org/10.1007/s10462-025-11403-7

Lucchi, N. (2023). *ChatGPT: A Case Study on Copyright Challenges for Generative AI Systems* (SSRN Scholarly Paper No. 4483390). Social Science Research Network. https://doi.org/10.2139/ssrn.4483390

Lucchi, N., Dornis, T. W., & Sierek, P. T. (2026). The three-step test in international copyright—A global framework for generative AI training. *American Business Law Journal*, *63*(3), 265–298. https://doi.org/10.1111/ablj.70022

Mahmud, S., & Roisah, K. (2023). The Concept of IPR Ownership in the Context of the Copyright Legal System and Indonesian Society's Cosmology. *International Journal of Social Science and Human Research*, *6*(08). https://doi.org/10.47191/ijsshr/v6-i8-12

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*,*1*(11),501–507. https://doi.org/10.1038/s42256-019-0114-4

Murtadho, N. A., Andreyani, R., Sakti, R. K., Rahmatullah, N., Salsabilah, Y. R., & Watae, S. (2024). Structure and Classification of Legal Norms: Institutional Challenges in Law Making. *Indonesian Journal of Administrative Law and Local Government*,*1*(01). https://doi.org/10.26740/ijalgov.v1i01.35853

Nasman, N., Astuti, P., & Perwitasari, D. (2024). Etika dan Pertanggungjawaban Penggunaan Artificial Intelengence di Indonesia. *Jurnal Hukum Lex Generalis*, *5*(10).https://ojs.rewangrencang.com/index.php/JHLG/article/view/622

Park, Y. J. (2013). Digital Literacy and Privacy Behavior Online. *Communication Research*, *40*(2),215–236. https://doi.org/10.1177/0093650211418338

Radanliev, P. (2025). AI Ethics: Integrating Transparency, Fairness, and Privacy in AI Development. *Applied Artificial Intelligence*, *39*(1).https://doi.org/10.1080/08839514.2025.2463722

Ramadhan, G. D. (2022). Perlindungan Hukum Terhadap Bahasa Pemrograman Dalam Undang-Undang Tentang Hak Cipta. *Lex Renaissance*,*7*(1),114–127. https://doi.org/10.20885/JLR.vol7.iss1.art9

Rees, P., & Hodgkinson, D. (2007). Binding Corporate Rules: A simpler clearer vision?. *Computer Law & Security Report*, *23*(4), 352–356. https://doi.org/10.1016/j.clsr.2007.03.004

Roisah, K., Rahayu, R., Asy'arie, M. A. H. A., Mitskaya, E., & Wahyudi, B. F. H. (2025). Empowering Cultural Communities in Protecting Traditional Expressions: Legal Challenges in The Digital Age. *Diponegoro Law Review*, *10*(2), 134–156. https://doi.org/10.14710/dilrev.10.2.2025.134-156

Rosadi, S. D., Yuniarti, S., & Fauzi, R. (2022). Protection of Data Privacy in the Era of Artificial Intelligence in the Financial Sector in Indonesia. *Journal of Central Banking Law and Institutions*, *1*(2), 353–366. https://doi.org/10.21098/jcli.v1i2.18

Sag, M. (2023). *Copyright Safety for Generative AI* (SSRN Scholarly Paper No. 4438593). Social Science Research Network. https://doi.org/10.2139/ssrn.4438593

Samiri, M., Kasim, M. A., Nonci, N., Ahmad, J., & Putra, M. A. A. (2025). Optimalisasi Pengelolaan Beban Kerja Hakim Dan Implikasinya Terhadap Integritas Dan Kualitas Peradilan. *Jurnal Pembangunan Hukum Indonesia*, *7*(2), 288–306. https://doi.org/10.14710/jphi.v7i2.108-127

Sapan, H. H., & Santiago, F. (2025). Ethics Court as an Answer to the Need for Enforcement of Ethical Norms in a Modern Democratic System. *International Journal of Social Service and Research*, *5*(7), 707–714. https://doi.org/10.46799/ijssr.v5i7.1254

Saparwadi, A., Priyono, F. J., & Setyowati, R. (2026). Determining Legal Status And Liability of Autonomous Artificial Intelligence System in Islamic Smart Contracts. *Diponegoro Law Review*, *11*(1), 38–67. https://doi.org/10.14710/dilrev.11.1.2026.38-67

Sarker, I. (2021). Deep Learning: A Comprehensive Overview on Techniques, Taxonomy, Applications and Research Directions. *SN Computer Science*, *2*. https://doi.org/10.1007/s42979-021-00815-1

Siau, K., & Wang, W. (2020). Artificial Intelligence (AI) Ethics: Ethics of AI and Ethical AI. *Journal of Database Management*, *31*, 74–

87.https://doi.org/10.4018/JDM.2020040105

Silviana, A., Elvani, R., & Priyono, E. A. (2025). The Digitalization of Land Titles as an Effort to Achieve Good Land Governance in the Framework of Reduced Inequalities. *IOP Conference Series: Earth and Environmental Science*,*1537*(1).https://doi.org/10.1088/1755-1315/1537/ 1/012031

Simanjuntak, E. N., Irmayani, D., & Nasution, F. A. (2024). Tinjauan Penerapan Kecerdasan Buatan Dalam Keamanan Jaringan: Tantangan Dan Prospek Masa Depan. *Jurnal Media Informatika*, *5*(2), 214–219. https://ejournal.sisfokomtek.org/index.php/jumin/article/view/3775

Simatupang, K. M. (2021). Tinjauan Yuridis Perlindungan Hak Cipta dalam Ranah Digital. *Jurnal Ilmiah Kebijakan Hukum*, *15*, 67–80. https://doi.org/10.30641/kebijakan.2021.V15.67-80

Spiekermann, S., Acquisti, A., Böhme, R., & Hui, K.-L. (2015). The challenges of personal data markets and privacy. *Electronic Markets*,*25*(2),161–167. https://doi.org/10.1007/s12525-015-0191-0

Tarikh, A., & Bahri, R. A. (2025). Analisis Perbandingan Hukum Perlindungan Data Antara Indonesia dan Uni Eropa: Tinjauan Kasus Cambridge Analytica. *Jurnal Penelitian Hukum Galunggung*, *2*(2), 1–16. https://doi.org/10.1234/jphgalunggung.v2i2.67

Turisno, B. E., Priyono, E. A., & Yunanto, Y. (2024). Strengthening Corporate Governance Through Legal Compliance and Ethical Codes of Conduct. *Jurnal Hukum Unissula*, *40*(2),214.https://doi.org/10.26532/jh.v40i2.40198

Tusquellas, N., Palau, R., & Santiago, R. (2024). Analysis of the potential of artificial intelligence for professional development and talent management: A systematic literature review. *International Journal of Information Management Data Insights*, *4*(2).https://doi.org/10.1016/j.jjimei.2024.100288

Wahyuni, R. A. E., & Turisno, B. E. (2019). Praktik Finansial Teknologi Ilegal Dalam Bentuk Pinjaman Online Ditinjau Dari Etika Bisnis. *Jurnal Pembangunan Hukum Indonesia*, *1*(3),379–391. https://doi.org/10.14710/jphi.v1i3.379-391

Wibowo, A., Alawiyah, W., & Azriadi, A. (2024). The importance of personal data protection in Indonesia's economic development. *Cogent Social Sciences*, *10*(1). https://doi.org/10.1080/23311886.2024.2306751

Yulianti, M., & Roisah, K. (2026). Komersialisasi Hak Cipta atas Konten Digital: Problematika Fair Use Doctrine dalam Perspektif Hukum Indonesia. *Journal of*

Creative Student Research, *4*(4), 65–82. https://doi.org/10.55606/jcsr-politama.v4i4.6482

Zelenka, N., Di Cara, N. H., Bennet, E., Clatworthy, P., Day, H., Garcia, I. K., Garcia, S. R., Hanschke, V. A., & Kuwertz, E. S. (2025). Data Hazards: An open-source vocabulary of ethical hazards for data-intensive projects. *Journal of Responsible Technology*, *21*. https://doi.org/10.1016/j.jrt.2025.100110

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Dong, Z., Du, Y., Yang, C., Chen, Y., Chen, Z., Jiang, J., Ren, R., Li, Y., Tang, X., Liu, Z., … Wen, J.-R. (2026). *A Survey of Large Language Models* (arXiv:2303.18223). arXiv.https://doi.org/10.48550/arXiv.2303.18223

PROCEEDINGS

Bakar, A. (2025, November 25). Law, Globalization, and Human Rights [Guest Lecturer]. In 2nd Postgraduate International Conference on Law, Technology, and Society. Yogyakarta: Faculty of Law, Universitas Islam Indonesia

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21* (pp.610–623). https://doi.org/10.1145/3442188.3445922

Budiyanti, R. T., Nandini, N., & Jati, S. P. (2025). Exploring Ethical and Legal Dimensions of Artificial Intelligence in Healthcare: Proposing a Regulatory Framework for Indonesia. *BIO Web of Conferences*, *193*, 00016. https://doi.org/10.1051/bioconf/202519300016

Du, R. (2026). Copyright Infringement Defenses for Generative AI Training: Comparing U.S. Fair Use, EU TDM Exceptions and the AI Act. In *Proceedings of the 2026 ICSSCS Symposium on Legal Studies and Social Governance (ICSSCS-LSSG 2026) 31* (pp.8–13). https://doi.org/10.70267/icsscs.lssg.2026v31.0813

Zhao, M., Agarwal, N., Basant, A., Gedik, B., Pan, S., Ozdal, M., Komuravelli, R., Pan, J., Bao, T., Lu, H., Narayanan, S., Langman, J., Wilfong, K., Rastogi, H., Wu, C.-J., Kozyrakis, C., & Pol, P. (2022). Understanding Data Storage and Ingestion for Large-Scale Deep Recommendation Model Training. In *Proceedings of the 49th Annual International Symposium on Computer Architecture* (pp.1042–1057). https://doi.org/10.1145/3470496.3533044

BOOKS

ASEAN Secretariat. (2024). *ASEAN Guide on AI Governance and Ethics*. Jakarta: ASEAN Secretariat

Bintoro, P., Ratnasari, R., Wihardjo, E., Putri, I. P., & Asari, A. (2024). *Pengantar Machine Learning*. Solok: PT. Mafy Media Literasi Indonesia

Kodner, I. J., Siegler, M., Freeman, D. M., & Choctaw, W. T. (2007). Ethical and Legal Considerations. In *The ASCRS Textbook of Colon and Rectal Surgery* (pp. 735–763). Springer, New York, NY. https://doi.org/10.1007/978-0-387-36374-5_53

Magnis-Suseno, F. (1993). *Etika Dasar: Masalah-Masalah Pokok Filsafat Moral* (Cet. 5). Yogyakarta: Penerbit Kanisius.

Marky, K. (2023). Data Collection Is Not Mostly Harmless: An Introduction to Privacy Theories and Basics. In N. Gerber, A. Stöver, & K. Marky (Eds.), *Human Factors in Privacy Research* (pp. 3–10). Springer International Publishing. https://doi.org/10.1007/978-3-031-28643-8_1

Nissenbaum, H. (2017). Protecting Privacy in an Information Age: The Problem of Privacy in Public. In Keith W Miller, Mariarosaria Taddeo (Eds.), *The Ethics of Information Technologies*. Routledge.

Roisah, K. (2015). *Konsep Hukum Hak Kekayaan Intelektual (Hki): Sejarah, Pengertian, Dan Filosofi Pengakuan Hki Dari Masa Ke Masa*. Malang: Setara Press.

Stahl, B. C., Schroeder, D., & Rodrigues, R. (2023). The Ethics of Artificial Intelligence: An Introduction. In B. C. Stahl, D. Schroeder, & R. Rodrigues (Eds.), *Ethics of Artificial Intelligence: Case Studies and Options for Addressing Ethical Challenges* (pp.1–7). Springer International Publishing. https://doi.org/10.1007/978-3-031-17040-9_1

ONLINE SOURCES

AI Ethics Communication Channel. (2020). The National Guidelines for AI Ethics. Retrieved from https://ai.kisdi.re.kr/eng/main/contents.do?menuNo=500011

Clark, B., & Kosinski, M. (2026). Apa itu Etika AI?. Retrieved from https://www.ibm.com/id-id/think/topics/ai-ethics

Compagnucci, M. C., Fenwick, M., & Haapio, H. (2024). Unlocking the Potential of Binding Corporate Rules (BCRs) in Health Data Transfers. Retrieved from https://doi.org/10.48550/arXiv.2407.21281

Cooper, A. F., & Grimmelmann, J. (2025). *The Files are in the Computer: On Copyright, Memorization, and Generative AI* (arXiv:2404.12590). Retrieved from https://doi.org/10.48550/arXiv.2404.12590

Datapribadi.id.(n.d). Naskah Akademik UU Pelindungan Data Pribadi. Retrieved from https://datapribadi.id/data-peraturan/naskah-akademik-ruu-pelindungan-data-pribadi/

European Comission. (2019). *Ethics guidelines for trustworthy AI | Shaping Europe's digital future*. Retrieved from https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

European Comission. (2021). Ethics by Design and Ethics of Use Approaches for Artificial Intelligence. Retrieved from https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/horizon/guidance/ethics-by-design-and-ethics-of-use-approaches-for-artificial-intelligence_he_en.pdf

G7 G20 Document Database. (2023). Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System. Retrieved from https://g7g20-documents.org/database/document/2023-g7-japan-leaders-leaders-annex-hiroshima-process-international-guiding-principles-for-organizations-developing-advanced-ai-system

Google Gemini. (2025). *GeminiAI - User Policy & Public Policy for Data Safety* [Computer software]. https://support.google.com/gemini/answer/14620100

Heryansyah, D. (2018). Sebuah Opini Despan Heryansyah tentang Etika dan Hukum. Retrieved from https://law.uii.ac.id/blog/2018/02/27/sebuah-opini-despan-heryansyah-tentang-etika-dan-hukum

JDIH BPK Database Peraturan. (2022). Undang-undang (UU) Nomor 27 Tahun 2022 Pelindungan Data Pribadi. Retrieved from https://peraturan.bpk.go.id/Details/229798/uu-no-27-tahun-2022

Korea Legislation Research Institute (KLRI). (2025). Framework Act on The Development of Artificial Intelligence And The Creation of A Foundation For Trust. Retrieved from https://elaw.klri.re.kr/eng_service/lawView.do?hseq=71019&lang=ENG

Meta Platforms. (2025). MetaAI - User Policy & Public Policy for Data Safety [Computer software]. Meta Platforms, Inc. https://transparency.meta.com/policies/

OECD Legal Instruments. (2024). Recommendation of the Council on Artificial Intelligence. Retrieved from https://legalinstruments.oecd.org/en/instruments/oecd-legal-0449

OpenAI. (2025). ChatGPT - User Policy & Public Policy for Data Safety (Version ChatGPT 5.6) [Indonesia]. Retrieved from https://chatgpt.com/share/6ae802c7-8698-83ec-ac5a-bda23e6f5ead

Regulations.ai. (2019). Japan - Human-Centric AI Principles; Social Principles of Human-

Centric AI. Retrieved from https://regulations.ai/regulations/RAI-JP-NA-SPHAXXX-2019

Search Engine Land. (2026). ChatGPT / OpenAI statistics: How many people use ChatGPT?. Retrieved from https://searchengineland.com/guide/chatgpt-stats

Tangf, H. (2025). Google Gemini User Statistics 2025: The Numbers Behind Its Explosive Growth. Retrieved from https://vertu.com/lifestyle/google-gemini-user-statistics-2025-the-numbers-behind-its-explosive-growth?srsltid=AYCX9vIuUOACNrmubcL8o4fwCRPs4OkryrSIHK8TxRnoi9XTbpBv7OzD

UNESCO. (2021). Recommendation on the Ethics of Artificial Intelligence. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000381137

Vaswani, J. (2025). Meta AI Users Statistics In 2026: Nearly 1 Billion Monthly Users and Growing. Retrieved from https://www.affiliatebooster.com/meta-ai-users/