# TRAINING AI MODELS ON COPYRIGHTED AND PERSONAL DATA: RECONCILING FAIR USE AND PRIVACY RIGHTS

Bedanta De, KIIT School of Law, KIIT University

**Page: 8869 - 8885**

## 1. ABSTRACT

Because of the fast advancements in generative AI, computers can now produce documents, pictures, music, and other content that resembles text written or made by people. A large portion of the information used to train these models is obtained from publicly available sources and contains a lot of copyrighted and personal data that people did not give permission for. It leads to legal debates regarding whether using such data is permitted by copyright law and privacy rules, especially since they are in the process of changing in India. Even though fair use and fair dealing are common arguments for AI training made by developers in the U.S. and India, they are coming under close inspection due to the privacy standards established by the Justice K.S. Puttaswamy v. Union of India which recognized the Right to privacy as a fundamental right under Article 21 of the Indian Constitution. Besides, India's latest Digital Personal Data Protection Act, 2023 places stricter rules on gathering, processing, and consent to the company's use of data, making it necessary to update AI systems. The paper examines the conflict between the use of copyrighted content in machine learning and the need to protect a person's privacy, mainly when information in the data is sensitive. Looking at the differences among India, America, and Europe, the study offers a review of the regulations and describes what changes should be made. It supports openness in where the data is obtained, open consent policies, and policies that equally guard progress and rights in the age of AI.

## 2. Introduction

Generative artificial intelligence is the new buzzword that has changed the landscape of creativity, productivity, and innovation in recent years. Such systems as the OpenAI's ChatGPT, Google Gemini, Anthropic Claude, and Stability AIS Stable Diffusion can produce human-like text, image, audio, and code data[1]. The machine models that enable such capabilities are trained on huge volumes of data including books, articles, art, websites, social media posts, and other publicly accessible or scraped data[2]. Nevertheless, the secrecy of the character and source of these data-files have made way to its strong legal discrimination especially concerning the aspect of copyright violations and unauthorized utilization of personal information[3].

Underlying this legal strife is the fact that AI models frequently consume copyrighted material and Personally Identifiable Information (PII) without the requisite consent or license. The copyright regime ascertains the legitimacy of such use in regards to whether it is fair use (in countries such as the U.S.) or fair dealing (as is the case under Indian law), and the generally accepted data protection laws challenge the validity of whether harvesting personal user data without their consent infringes on the individual, in terms of their right to privacy[4]. In India, this has been the most pertinent debate with the Supreme Court having made the landmark this year in the case of Justice K.S. Puttaswamy v. Union of India, where the right of informational privacy was again held to be inherent in Article 21 of the Constitution[5], the latest statute regarding the very securing of such the provision being the Digital Personal Data Protection Act 2023, which aims at putting in practical effect the principle of consent-based data processing[6].

---

[1] See generally OpenAI, https://openai.com; Google, *Introducing Gemini: Our Most Capable AI Model*, https://blog.google/technology/ai/google-gemini-ai; Anthropic, https://www.anthropic.com; Stability AI, https://stability.ai.

[2] European Commission, *Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*, COM (2021) 206 final (Apr. 21, 2021), https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206.

[3] See U.S. Copyright Office, *Artificial Intelligence and Copyright*, 88 Fed. Reg. 19,898 (Apr. 4, 2023), https://www.federalregister.gov/documents/2023/04/04/2023-06958/artificial-intelligence-and-copyright.

[4] Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), 2016 O.J. (L 119) 1 (EU).

[5] *Justice K.S. Puttaswamy v. Union of India*, (2017) 10 SCC 1 (India).

[6] The Digital Personal Data Protection Act, No. 22 of 2023, Acts of Parliament, 2023 (India), https://www.meity.gov.in/content/digital-personal-data-protection-act-2023.

At the same time, in different jurisdictions, several lawsuits are created against big names in AI development. As an example, writers and musicians have claimed that their work that is under copyright has been deployed unauthorized to train AI that can now compete with human creations. In the privacy domain, there are concerns with law scholars and technologists about the harvesting of facial images, samples of voice recordings and the history of interactions with an AI company that could be used in identity theft, profiling or by causing reputational harm[7].

The purpose of this paper is to be able to critically analyse the legal and ethical paradox between the utilization of copyrighted and individual data to teach AI models and safeguards provided through the copyright and privacy laws. It is most centred within India, and makes comparative observations concerning the fair use doctrine in the United States, the General Data Protection Regulation (GDPR) in the European Union, and other international case law. The goal is to de-mutualize and have a balance in which technological innovation becomes a possibility and yet constitutional and statutory rights, especially authorship, consent, and dignity are allowed to prevail.

[7] Kashmir Hill, *Clearview AI's Facial Recognition App Is Identifying People at a Distance*, N.Y. Times (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.

## 3. Understanding AI Training Models and Their Data Sources

Generative Artificial Intelligence depends on machine learning algorithms, particularly large language models (LLMs) and generative adversarial networks (GANs), which cannot learn patterns, styles, and associations without access to vast datasets. The effectiveness and sophistication of these systems are heavily influenced by the comprehensiveness and diversity of the data on which they are trained[1]. In practice, this training data often includes publicly available information as well as copyrighted materials such as news articles, artwork, code repositories, personal blogs, and social media content scraped from the internet[2].

Generative AI functions differently from traditional software. While traditional software operates based on pre-written rules and logic, generative AI learns and generates outputs based on patterns in data. For instance, models like OpenAI's GPT-4 or Stability AI's Stable Diffusion are trained on internet-scale datasets, which may include unlicensed data scraped from websites such as Wikipedia, Reddit, Stack Overflow, Instagram, DeviantArt, and others[3]. While such practices enhance the realism and utility of generated content, they raise legal and ethical concerns regarding whether the collection and use of such data are authorized[4]. Much of this scraped data may be copyrighted or may include personally identifiable information (PII), which is subject to privacy regulations such as the GDPR and CCPA[5].

---

[1] Alex Hanna & Emily Denton, *Towards a Critical Data Practice: Reflections on the AI Dataset Landscape*, 3 Proc. ACM on Fairness, Accountability, & Transparency 1 (2020).
[2] Irene Solaiman et al., *Release Strategies and the Social Impacts of Language Models*, 3 FAccT 146, 148 (2021).
[3] Kashmir Hill, *Clearview AI's Facial Recognition App Is Identifying People at a Distance*, N.Y. Times (Jan. 18, 2020), https://www.nytimes.com/2020/01/18/technology/clearview-privacy-facial-recognition.html.
[4] Pamela Samuelson, Big Data, Artificial Intelligence, and Copyright, 13 Harv. J.L. & Tech. 69, 73 (2020).
[5] *General Data Protection Regulation (GDPR): Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, 2016 O.J. (L 119) 1.*

The materials of training the model fall into the following three types:

- Public Domain Material: these can be texts and media whose copyrights have either passed out of copyright or have been licensed to be used freely[6].
- Licensed Data Sets: This is data that was created in a specific way, and is legally acquired through contracts or open licenses including Creative Commons[7].
- Unlicensed or Scraped Data: Data retrieved through Web-based crawlers which could have a copyright or privacy and collected without the knowing permission of the authors or the subjects[8].

One of the key issues in AI development is the lack of transparency in the documentation of training datasets. For example, OpenAI has been criticized for not disclosing the exact composition of the datasets used to train its GPT models, citing commercial interests. Similarly, Meta's LLaMA series and the LAION-5B dataset developed by Stability AI have faced criticism for including copyrighted or personal images without clear user consent[9]. This lack of transparency hinders accountability and makes it difficult to assess potential violations of legal rights.

Moreover, the use of personal data in training sets introduces significant privacy risks. Even if such data is later anonymized or transformed, generative AI models may reproduce or "hallucinate" real names, phone numbers, or addresses unintentionally. This directly impacts the right to informational privacy[10].

As AI systems become increasingly adopted in both commercial and governmental contexts, it is essential to scrutinize the nature and sources of training data. Clearly defining the boundary between fairly and unfairly accessed data is fundamental not only to comply with intellectual property laws but also to protect the rights to privacy and dignity, which are essential in any democratic society[11].

---

[6] Mark A. Lemley, *IP in a World Without Scarcity*, 90 N.Y.U. L. Rev. 460, 478–79 (2015).
[7] Creative Commons, About the Licenses, https://creativecommons.org/licenses/ (last visited June 8, 2025).
[8] Tim Hwang**,** *Data Protection and Machine Learning: Models, Norms and Legal Challenges*, 5 Colo. Tech. L.J. 31, 36 (2021).
[9] James Vincent, *Why Meta's Latest AI Model, LLaMA, Is Making Waves*, The Verge), https://www.theverge.com/2023/2/24/23613266/meta-ai-language-model-llama-release-research.
[10]Justice K.S. Puttaswamy v. Union of India, (2017) 10 S.C.C. 1 (India).
[11]Shreya Srinivasan, *Fairness and Accountability in AI: A Human Rights Perspective*, 9 Indian J.L. & Tech. 122, 128 (2023).

## 4. Copyright Concerns in AI Training

One of the most urgent legal issues in the current intellectual property landscape is the use of copyrighted material in training AI models. Legal copyright protection assures authors the exclusive rights to reproduce, adapt, and publish their works[1]. However, generative AI models such as GPT-4, Midjourney, and DALL·E require massive datasets to train, often including millions of books, articles, images, and code, frequently without the knowledge or authorization of the original copyright owners[2]. This practice raises critical legal questions: Is the use of copyrighted material for AI training illegal? Can developers claim immunity under doctrines like *fair use* or *fair dealing*?

In India, the Copyright Act, 1957 provides authors a bundle of exclusive rights, such as the right to reproduce and the right to communicate works to the public[3]. It also contains a fair dealing exception, allowing limited use of copyrighted works for purposes such as private use, criticism, review, and reporting of current events. However, the Act does not explicitly address the automated and large-scale ingestion of copyrighted content for commercial purposes, such as AI model development. This legislative ambiguity places AI developers at risk of infringement, particularly if their use fails to meet the threshold of being transformative or does not involve proper attribution[4].

In contrast, the United States adopts a broader fair use doctrine under Section 107 of the U.S. Copyright Act, which allows the use of copyrighted works without permission when used for purposes such as research, education, or commentary, provided it passes a four-factor test[5]:

- The purpose and character of the use (e.g., whether it is transformative or commercial),
- The nature of the copyrighted work,
- The amount and substantiality of the portion used, and
- The effect on the potential market for the original.

---

[1] The Copyright Act, No. 14 of 1957, INDIA CODE (1957).
[2] Andres Guadamuz, *Artificial Intelligence and Copyright*, 39 W. New Eng. L. Rev. 103, 105 (2017).
[3] Indian Copyright Act, 1957, §§ 14–16
[4] Arul George Scaria, *AI and Copyright Law in India: A Critical Assessment*, 11 Indian J.L. & Tech. 93, 100–01 (2023).
[5] 17 U.S.C. § 107 (2021).

In Authors Guild v. Google, Inc., the U.S. Court of Appeals held that Google's digitization of books to create a searchable index qualified as fair use, emphasizing its transformative nature and minimal market harm[6]. However, whether the training of AI models constitutes a similarly transformative use remains unresolved and is likely to be tested in court.

Recent lawsuits reflect the growing unease among copyright holders. Visual artists have alleged that AI models trained on copyrighted images without permission can produce derivative works that directly compete in the market[7]. Similarly, authors have argued that large language models like those developed by OpenAI are capable of reproducing copyrighted text verbatim, thereby infringing the right of reproduction[8]. These legal actions signal an impending judicial reckoning, which may set precedents on whether AI training constitutes fair use or requires a licensing framework.

Scholars and critics argue that unless AI training serves a socially beneficial or significantly transformative purpose, it may not satisfy the fair use test[9]. Furthermore, AI outputs often closely mimic the style of specific artists or reproduce substantial portions of existing texts, blurring the line between learning and copying. In such cases, AI-generated outputs may be classified as unauthorized derivative works, a clear infringement under both Indian and international copyright laws[10].

In India, this remains a legal grey area due to a lack of judicial interpretation or statutory guidance. While the Indian Copyright Office has not yet issued formal directives on AI training, stakeholders are increasingly calling for regulatory clarity, particularly around attribution, compensation, and licensing. The challenge lies in striking a balance between fostering technological innovation and safeguarding the moral and economic rights of authors, especially as AI-generated content becomes nearly indistinguishable from that created by humans[11].

---

[6]Authors Guild v. Google, Inc., 804 F.3d 202 (2d Cir. 2015).
[7]James Vincent, *Artists File Class-Action Lawsuit Against AI Image Generators*, The Verge (Jan. 17, 2023), https://www.theverge.com/2023/1/17/23558928.
[8]Joe Mullin, *Authors Sue OpenAI, Say ChatGPT Copies Their Books*, Ars Technica (June 29, 2023), https://arstechnica.com/tech-policy/2023/06/authors-sue-openai-say-chatgpt-copies-their-books/
[9] Shreya Srinivasan, Fairness and Accountability in AI: A Human Rights Perspective, 9 Indian J.L. & Tech. 122, 128 (2023).
[10] Dina Srinivasan, The Great Internet Brand Rip-Off: How the Internet Is Being Used to Appropriate Brand Value, 2 Hastings Sci. & Tech. L.J. 49, 73 (2010).
[11] Pankhuri Agarwal**,** *India's Copyright Framework for Artificial Intelligence: Challenges and the Way Forward*, 15 NALSAR Student L. Rev. 210, 214 (2023).

## 5. Privacy Rights and Data Protection

Although, in the legal context, copyright issues are the primary focus of objections to AI training, individual privacy is another critical concern that must be addressed[1]. Since AI models are trained on large datasets that may contain personal user information, there is an ongoing concern about the unauthorized processing, collection, and potential disclosure of sensitive data[2]. These concerns strike at the core of informational privacy, i.e., the right of individuals to control what happens to their personal data, with whom it is shared, and for what purpose[3].

In India, the right to privacy has been declared a fundamental right under Article 21 of the Constitution[4]. This encompasses the right to informational privacy, which protects individuals against the unregulated and unjust collection or surveillance of their data. AI models often crawl the internet, including social media sites, blogs, forums, and other publicly accessible sources[5]. One of the central concerns is that personally identifiable information (PII), such as names, faces, locations, and even behavioural data, may be scraped and incorporated into training datasets[6].

This becomes especially problematic when, during inference, AI models produce outputs that unintentionally reveal confidential information about real individuals[7]. For example, chatbots or image generators may replicate personal characteristics or produce outputs that closely resemble a person's digital footprint[8]. These results not only infringe on privacy but may also lead to reputational harm, identity theft, or psychological distress[9].

[1]Pankhuri Agarwal, India's Copyright Framework for Artificial Intelligence: Challenges and the Way Forward, 15 NALSAR Student L. Rev. 210, 214 (2023).
[2]Sonia Livingstone, Rethinking the Rights of Children for the Digital Age, 29 L. & Pol'y 431, 436 (2007).
[3]Daniel J. Solove, *Understanding Privacy* 24–27 (2008).
[4]*Justice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1.
[5]Jenna Burrell, How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms, 3 *Big Data & Soc'y* 1, 3–5 (2016).
[6]Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE Symp. on Sec. & Privacy 111.
[7]Reuben Binns et al., *'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions*, in *CHI '18 Proceedings* 377.
[8]Sandra Wachter & Brent Mittelstadt, *A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI*, 2019 Colum. Bus. L. Rev. 494, 499–500.
[9]Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* 71–73 (2018).

The recent enactment of the Digital Personal Data Protection Act**,** 2023 (DPDPA) in India marks an important step toward developing a comprehensive data protection regime[10]. The Act emphasizes principles such as purpose limitation, data minimization, and informed consent. Under this law, any organization that collects or processes personal data must obtain explicit consent from the data principal and use the data strictly for the stated purpose. These requirements are often at odds with the practices of many AI firms, which typically gather data by indiscriminately crawling websites, frequently without the awareness or consent of users[11]. At the international level, data protection compliance is more robust. The General Data Protection Regulation **(**GDPR**)** of the European Union, for instance, establishes rights such as the right to be forgotten, the right to data portability, and mandates data protection impact assessments[12]. These rules require that AI developers, particularly those working with international datasets, have a lawful basis for processing personal data, especially when such data is used to train AI models. Despite these legal advancements, a significant regulatory gap remains concerning generative AI. In India, as in many other jurisdictions, there are currently no specific provisions governing the training of AI models using personal data[13]. This legal ambiguity creates uncertainty for developers regarding the permissible scope of data usage. The lack of dataset transparency and the absence of strong enforcement mechanisms further elevate the risk of mass privacy violations[14].

Therefore, AI system design must not only comply with legal frameworks but also be guided by ethical considerations[15]. This includes conducting audits of training datasets, removing sensitive information, and implementing mechanisms to prevent or filter out personal data during model training[21]. As AI continues to evolve, ensuring that innovation does not come at the cost of individual dignity and autonomy will remain a critical challenge for lawmakers, technologists, and society at large.

---

[10]The Digital Personal Data Protection Act, No. 22 of 2023, Acts of Parliament, 2023 (India).
[11]Pankhuri Agarwal, supra note 1, at 219.
[12]Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data (General Data Protection Regulation), arts. 17, 20, 2016 O.J. (L 119) 1.
[13]Karan Saini, *Generative AI and Indian Law: An Emerging Legal Frontier*, 18 Indian J.L. & Tech**.** 76, 85 (2024).
[14]Malavika Raghavan**,** *India's Data Protection Law: Challenges and Opportunities*, 13 NUJS L**.** Rev**.** 1, 10–12 (2023).
[15]Luciano Floridi & Josh Cowls, *A Unified Framework of Five Principles for AI in Society*, 5 Harv. Data Sci. Rev. 1, 3 (2020).
[16]Andrew D. Selbst et al., Fairness and Abstraction in Sociotechnical Systems, in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 59 (2019).

# 6. Reconciling Fair Use and Privacy: A Legal Dilemma

The combined application of the doctrines of fair use (or fair dealing) and the right to privacy creates a complex and evolving legal conundrum in the context of artificial intelligence training models[1]. Both principles are intended to safeguard distinct yet equally vital public interests, fair use facilitates innovation, knowledge sharing, and creativity, while the right to privacy upholds individual autonomy, dignity, and control over personal data[2]. However, in the rapidly growing domain of generative AI, these doctrines are increasingly being juxtaposed and brought into tension[3]. This raises a pivotal question for scholars, developers, and lawmakers alike: Can the use of data for AI model training be considered fair and lawful if it encroaches upon the privacy rights of individuals whose data is involved?[4]

Copyright law in many jurisdictions, including the United States, United Kingdom, and India, permits the fair use or fair dealing of copyrighted content[5]. This refers to the limited use of such material without the explicit consent of the copyright holder, typically for socially valuable purposes such as education, commentary, research, or criticism. In the context of AI, developers and technology firms frequently argue that the training of machine learning models on large volumes of data should be interpreted as a form of research or innovation and, as such, should fall within the protective ambit of fair use[6]. They contend that this exemption is essential for technological progress and for maximizing the capabilities of AI systems.

However, this argument tends to prioritize innovation over individual rights and often fails to account for informational privacy, a legal and ethical concept that emphasizes a person's right to govern the collection, use, and dissemination of their personal information[7].

---

[1]Pankhuri Agarwal**,** *India's Copyright Framework for Artificial Intelligence: Challenges and the Way Forward*, 15 NALSAR Student L. Rev**.** 210, 213 (2023).
[2]Daniel J. Solove, *Understanding Privacy* 6–8 (2008).
[3]Andrew D. Selbst et al., Fairness and Abstraction in Sociotechnical Systems, in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 59 (2019).
[4]Woodrow Hartzog, *Privacy's Blueprint: The Battle to Control the Design of New Technologies* 47–48 (2018).
[5]Copyright Act, 17 U.S.C. § 107 (U.S.); Copyright, Designs and Patents Act 1988, c. 48 (UK); Copyright Act, No. 14 of 1957, § 52 (India).
[6]Rebecca Giblin & Cory Doctorow, *Chokepoint Capitalism* 193–95 (2022)
[7]Solove, supra note 2, at 37–41

AI training datasets frequently contain not just publicly available data but also sensitive and personally identifiable information (PII), such as names, faces, addresses, conversations, preferences, and behavioral profiles[8]. In many cases, such data is scraped from the internet without the knowledge or consent of the individuals concerned[9].

AI systems generally rely on data scraping techniques, extracting content from a wide range of public web platforms, including social media, blogs, forums, and open databases. While developers often claim that public availability implies consent or legal permissibility, the reality is more nuanced[10]. The mere accessibility of information on the internet does not imply a waiver of privacy rights. Personal data that is publicly visible, whether intentionally shared or not, still carries with it legal protections under data privacy frameworks[11]. Therefore, what may be considered permissible under the principles of copyright (e.g., using text or images for machine training) may simultaneously constitute a breach of privacy when the data involved is personal in nature and obtained without consent[12].

This legal conflict becomes especially prominent in jurisdictions like India, where the right to privacy has been constitutionally enshrined as a fundamental right under Article 21 following the landmark decision of K.S. Puttaswamy v. Union of India in 2017[13]. Indian data protection law, as embodied in the Digital Personal Data Protection Act, 2023 (DPDPA), places strong emphasis on consent, purpose limitation, data minimization, and the protection of personal data[14]. Under this regulatory scheme, AI developers using datasets that include identifiable personal data, such as names, images, or location information, must obtain informed consent from the data principal. Even if the use of such data is deemed transformative or non-commercial, it may not pass legal muster if consent was not obtained, thereby rendering a fair use defense inadequate in such circumstances[15].

---

[8] Arvind Narayanan & Vitaly Shmatikov, *Robust De-anonymization of Large Sparse Datasets*, 2008 IEEE Symp. on Sec. & Privacy 111.
[9] Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 Big Data & Soc'y 1, 3–5 (2016).
[10] Sandra Wachter, Brent Mittelstadt & Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, 7 Int'l Data Privacy L. 76, 80–82 (2017).
[11] General Data Protection Regulation, Regulation (EU) 2016/679, arts. 4, 6, 9.
[12] Malavika Raghavan, India's Data Protection Law: Challenges and Opportunities, 13 NUJS L. Rev. 1, 8–10 (2023).
[13] *Justice K.S. Puttaswamy (Retd.) v. Union of India*, (2017) 10 SCC 1.
[14] Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).
[15] Raghavan, supra note 16, at 12.

At the same time, it must be acknowledged that existing privacy regulations, both in India and internationally, are not without gaps[16]. One significant limitation is that most legal frameworks focus predominantly on identifiable personal data, and tend to leave anonymized or aggregated data outside the scope of stringent regulation. AI developers often exploit this loophole by arguing that training data has been anonymized and thus no longer poses a risk to individual privacy. However, with advancements in re-identification techniques and inferential modeling, even anonymized datasets can be reverse-engineered to reveal sensitive patterns, identities, or behaviors[17]. This creates a grey zone in law, where datasets are technically anonymized but still capable of generating outputs that reflect, replicate, or approximate real individuals[18].

Reconciling the twin doctrines of fair use and privacy thus requires a comprehensive legal and policy response, tailored specifically to the unique challenges of artificial intelligence[19]. First and foremost, there is a pressing need to redefine the contours of fair use and fair dealing in the digital and AI era[20]. Legislators should provide clearer guidance on whether and how large-scale, automated data ingestion by AI systems can qualify as fair use. This would involve consideration of factors such as the nature of the data, the purpose of use, the presence or absence of consent, and the degree of transformation involved[21].

Secondly, privacy regulations must evolve to accommodate the risks posed by machine learning[22]. Laws must account for the possibilities of re-identification, inference-based profiling, and the generation of synthetic data that still mimics real human attributes[23]. Provisions must be introduced to ensure transparency in data sourcing, allow for audits of training datasets, and impose strict penalties for non-consensual or unethical data use[24]. A duty to document datasets, outlining the nature, origin, and processing rationale, should be mandated, especially for models intended for public interaction or deployment at scale[25].

---

[16]Karan Saini, *Generative AI and Indian Law: An Emerging Legal Frontier*, 18 Indian J.L. & Tech. 76, 83 (2024).
[17]Narayanan & Shmatikov, supra note 10.
[18]Wachter & Mittelstadt, supra note 13, at 84–85.
[19]Floridi & Cowls, *A Unified Framework of Five Principles for AI in Society*, 5 Harv. Data Sci. Rev. 1, 3 (2020).
[20]Giblin & Doctorow, supra note 7, at 201.
[21]Hartzog, supra note 4, at 71.
[22]Solove, supra note 2, at 105–106.
[23] Wachter et al., supra note 13, at 90.
[24] Raghavan, supra note 16, at 16–17.
[25] Selbst et al., supra note 3, at 65.

Most critically, there is a need to develop a "consent-aware" model of fair use[26]. Under such a framework, the law would require not just an analysis of the purpose and extent of data use but also an evaluation of whether the use was accompanied by informed consent and meaningful privacy protections. This holistic approach would ensure that innovation proceeds within ethical and legal boundaries, and that AI technologies are developed in a way that respects both the creative commons and individual rights[27].

In conclusion, while fair use and privacy each serve indispensable roles in the legal ecosystem, their intersection in the context of AI training calls for a new paradigm, one that recognizes their overlaps and conflicts and seeks balance through legislative clarity, ethical design choices, and robust enforcement[28]. Only then can we achieve a future where artificial intelligence is both innovative and responsible.

[26] Floridi & Cowls, supra note 26, at 6.
[27] Hartzog, supra note 4, at 112.
[28] Agarwal, supra note 1, at 220.

## 7. Proposed Legal and Policy Frameworks

With the growing legal tension between copyright protection, fair use, and privacy rights in the context of artificial intelligence, it has become evident that existing frameworks are ill-equipped to address the complexities of generative AI[1]. As AI systems increasingly rely on large-scale data ingestion, often without transparency or consent, there is an urgent need for legal and policy reforms that strike a balance between innovation and the rights of individuals and content creators[2].

One of the foremost issues is the opacity surrounding AI training datasets[3]. Many companies refuse to disclose whether their models are trained on copyrighted or personal data, operating in secrecy and avoiding accountability[4]. A regulatory mandate requiring public disclosure of datasets, particularly when personal or protected content is involved, would enhance transparency and enable affected individuals to challenge unlawful usage. Such disclosure is essential for fostering accountability and ensuring compliance with both copyright and privacy standards[5].

Second, privacy legislation must evolve to address the realities of AI development[6]. Most data protection laws, including India's Digital Personal Data Protection Act, 2023, focus on identifiable personal information within structured databases[7]. However, AI training typically involves vast volumes of unstructured data, and even anonymized datasets pose risks of re-identification. Privacy frameworks must, therefore, incorporate explicit provisions governing AI training, including safeguards such as data minimization, purpose limitation, and mandatory anonymization processes. These principles are essential for protecting informational privacy in machine learning contexts.

---

[1]Pankhuri Agarwal, *India's Copyright Framework for Artificial Intelligence: Challenges and the Way Forward*, 15 NALSAR Student L. Rev. 210, 214 (2023).
[2]Andrew D. Selbst et al., Fairness and Abstraction in Sociotechnical Systems, in *Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency* 59, 61 (2019).
[3]Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, 3 Big Data & Soc'y 1, 3–5 (2016).
[4]Karan Saini, *Generative AI and Indian Law: An Emerging Legal Frontier*, 18 Indian J.L. & Tech. 76, 83–84 (2024).
[5]Sandra Wachter, Brent Mittelstadt & Luciano Floridi, Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation, 7 INT'L DATA PRIVACY L. 76, 78–80 (2017).
[6]Malavika Raghavan, *India's Data Protection Law: Challenges and Opportunities*, 13 NUJS L. Rev. 1, 10–12 (2023).
[7]The Digital Personal Data Protection Act, No. 22 of 2023, Acts of Parliament, 2023 (India).

Equally important is the development of a consent-based framework across both copyright and data protection regimes[8]. In the copyright domain, this could include mechanisms like collective licensing systems or opt-out registries where creators can deny the use of their work in AI training. For personal data, a robust informed consent mechanism must be embedded in AI development processes, reinforcing individuals' rights to control their digital identity and personal information, even when such data is publicly available online[9].

The global nature of AI systems also demands international cooperation[10]. Institutions such as the World Intellectual Property Organization (WIPO), UNESCO, and the OECD have initiated discussions on ethical and legal frameworks for AI governance[11]. Given its growing influence in the tech and legal sectors, India should take a leadership role in shaping international norms[12]. Harmonizing AI-related laws across jurisdictions is vital, especially considering that data flows and AI applications often cross national borders, creating enforcement and legal uncertainty.

From a regulatory perspective, the establishment of independent AI oversight bodies or ethics boards is imperative[13]. These institutions should be empowered to audit training datasets, address complaints, and impose penalties for violations. They should also focus on raising awareness about the legal rights of individuals and creators whose data may be used in AI systems. Such regulatory bodies will be critical to ensuring both preventive and remedial measures in cases of infringement[14].

Finally, there is a need to promote research and civil society engagement. Think tanks, academic institutions, and advocacy organizations must be encouraged to conduct impact assessments of generative AI systems, particularly with regard to privacy, intellectual property, and democratic values. These assessments can inform evidence-based policymaking and support the development of a rights-respecting AI ecosystem[15].

---

[8]Rebecca Giblin & Cory Doctorow, *Chokepoint Capitalism* 193–95 (2022).
[9]Daniel J. Solove, *Understanding Privacy* 24–27 (2008).
[10]Floridi & Cowls, *A Unified Framework of Five Principles for AI in Society*, 5 Harv. Data Sci. Rev. 1, 4–5 (2020).
[11]OECD, *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (2019); WIPO, *Revised Issues Paper on Intellectual Property Policy and Artificial Intelligence* (2020); UNESCO, *Recommendation on the Ethics of Artificial Intelligence* (2021).
[12]Karan Saini, supra note 4, at 88.
[13]Selbst et al., supra note 2, at 65.
[14]Agarwal, supra note 1, at 219.
[15]Floridi & Cowls, supra note 15, at 7.

## 8. Conclusion

The blistering pace of development in generative artificial intelligence has brought about transformative changes in the creation, processing, and consumption of information[1]. While these technological advancements hold significant potential for fostering innovation, efficiency, and creativity, they simultaneously pose complex legal and ethical challenges, particularly in the domains of copyright protection, fair use, and the right to privacy[2]. The current legal frameworks appear increasingly inadequate, as AI systems rely heavily on large volumes of data, much of which is collected without explicit permission or clear legal entitlement[3].

This paper explores the intersection between the legal doctrines of fair use and the right to privacy in the context of AI training, and how these two rights, though rooted in different objectives, can come into conflict[4]. Copyright law aims to protect the economic and moral rights of content creators, ensuring control over the use of their intellectual property. Conversely, privacy law is designed to safeguard individuals from unauthorized access, use, or disclosure of their personal information[5]. The training of AI models on datasets that may contain both copyrighted and personal data reveals a significant legal grey area, where current laws provide neither clear guidelines nor effective remedies.

This tension is particularly pronounced in India, where the right to privacy has been constitutionally recognized under Article 21, and where the recently enacted Digital Personal Data Protection Act, 2023, seeks to regulate data use at the national level[6]. Nevertheless, neither India's copyright framework nor its data protection regime offers a comprehensive solution tailored to the complexities introduced by generative AI[7]. This lack of clarity results in legal uncertainty for developers, confusion among users, and potential harm to individuals whose data may be unknowingly processed.

---

[1] Organisation for Economic Co-operation and Development (OECD), *Recommendation of the Council on Artificial Intelligence*, OECD/LEGAL/0449 (May 22, 2019).
[2] Pankhuri Agarwal**,** *India's Copyright Framework for Artificial Intelligence: Challenges and the Way Forward*, 15 NALSAR Student L. Rev**.** 210, 212–13 (2023).
[3]Jenna Burrell, How the Machine "Thinks": Understanding Opacity in Machine Learning Algorithms, 3 *Big Data & Soc'y* 1, 3–4 (2016).
[4]Andrew D. Selbst et al., *Fairness and Abstraction in Sociotechnical Systems*, 81 Proc. of the Conf. on Fairness, Accountability, and Transparency 59, 61 (2019).
[5]Daniel J. Solove, *Understanding Privacy* 24–25 (2008).
[6]Digital Personal Data Protection Act, 2023, No. 22, Acts of Parliament, 2023 (India).
[7]Karan Saini**,** *Generative AI and Indian Law: An Emerging Legal Frontier*, 18 Indian J**.**L. & Tech**.** 76, 81–82 (2024).

To address this regulatory gap, a harmonized legal framework is necessary, one that upholds the core principles of transparency, consent, accountability, and fairness. Legislatively, such reform should clarify the scope and limitations of fair dealing exceptions as they apply to AI-generated content, establish explicit consent requirements for the use of personal data in training models, and provide accessible avenues for redress and enforcement. Beyond statutory reform, it is equally essential to promote ethical AI design, raise stakeholder awareness, and contribute to the formulation of international standards that can guide AI development across jurisdictions.

The fundamental point is that the future of AI must not be built at the expense of basic human rights. A hybrid, rights-sensitive regulatory approach is essential to ensure that generative AI evolves in a way that respects the creative contributions of authors while simultaneously protecting individual dignity and autonomy. Such a model is not only legally and ethically desirable but is also essential for building a sustainable and inclusive digital future.