TRAINING IN THE DARK – WHY COPYRIGHT LAW NEEDS DATASET TRANSPARENCY IN THE AGE OF GENERATIVE AI

Noor Bansal and Mishti Bansal[1]

Abstract

Before artificial intelligence can dream, it must first consume, and what it consumes may belong to someone else. This paper analyses the copyright implications arising at the training stage of Artificial Intelligence models, where datasets may include copyrighted content depending on their composition and licensing, and examines how these issues are interpreted across various jurisdictions, with a focus on Ghibli Studios as a Case Study. By examining the laws of the various countries, the article sheds light on how different jurisdictions define and interpret 'authorship' and the legality of data mining practices. While several nations have taken meaningful steps and initiatives to address the issue of AI-generated works, significant ambiguity persists especially regarding the extent of infringement, fair use, and the protection of original creators, as courts are yet to definitively resolve these issues, especially in relation to the training stage of AI models. This article will further go on to explore how in the case of rapid developments in generative AI, the questions of legal compliance have moved from theoretical to urgent. In conclusion, the article critiques the current legislative landscape and argues that regulatory transparency at the training stage is essential.

I. INTRODUCTION

Alan M. Turning defines Artificial Intelligence as "the development of computer systems capable of performing tasks that require human-like intelligence."[2] Since the debut of ChatGPT in late 2022, policymakers around the world have been engaged and captivated by generative AI, i.e., the models which can generate various types of content, including text, images, audio, and video.[3] Generative AI is seen as a groundbreaking technology with the potential to transform industries. Many experts believe it could contribute trillions of dollars to the global economy in the coming decade.[4] Although there have been AI models that are somewhat "generative" for decades, the latest generation of well-known generative AI models, including OpenAI's GPT series and Stability AI's image creator, relies on a specialized category of machine learning called 'deep learning.' Similar to other machine learning approaches, deep learning employs 'neural networks,' which are interconnected nodes designed to mimic the architecture of the human brain. As information moves through these deep layers, it undergoes processing at various levels of complexity. The initial layers focus on detecting simple patterns, while the later layers build upon this groundwork to identify more complex patterns. This hierarchical approach

[1] Authors are Second-Year B.A. LL.B. (Hons.) Students at National Law University, Jodhpur.
[2] A.M. Turing, 'Lecture to the London Mathematical Society' in DC Ince (ed), Mechanical Intelligence: Collected Works of A M Turing (North Holland 1992).
[3] Adam Zewe, 'Explained: Generative AI' (MIT Schwarzman College of Computing, 9 November 2023) <https://computing.mit.edu/news/explained-generative-ai/> accessed 14 January 2026.
[4] Douglas Broom, 'How generative AI could add trillions to the global economy' (World Economic Forum, 14 July 2023).

allows AI models to effectively learn from vast amounts of data.[5] Data is a crucial component of the process, and is required to train models to understand and generate new text. Applications include translation, information extraction, and question answering.[6] The deep learning-based generative AI model, identifies the patterns and principles during training which are utilized to produce new content.

The machine learning algorithms also utilise statistics to find insights hidden in massive amounts of data.[7] A machine learning algorithm identifies patterns from past data and produces a model that encapsulates these patterns. The process begins with defining the problem (i.e., determining what the model should predict). "*An AI developer must collect data, prepare the data (including labelling or annotating), choose a model architecture; train the model with the annotated data (training data) using a machine learning algorithm, and test the model with new data (testing data).*"[8] For instance, a developer aiming to teach a computer to identify a surgeon's hand movements during surgery would begin with a model previously trained on ImageNet, which includes over 14 million images of everyday objects such as cats, dogs, and cars. In addition to the promised opportunities, however, generative AI also presents significant challenges, such as replicating or training data on copyrighted content and endangering the livelihood of human workers (especially in the creative industries).[9] Copyright law protects the originality and creativity of human authors.[10] In India, the protection of copyright is governed by The Copyright Act, 1957, while at the international level, the Berne convention, TRIPS Agreement, and the WIPO Copyright Treaty govern it. However, this established legislative policy has been disrupted by the emergence of generative AI and its involvement in creating diverse forms of work, such as artistic or literary works.[11] The main issue is that AI-generated content often uses the largest training datasets comprising millions of texts, images, audio samples, or other forms of content. The training stage input is often scraped from the internet, which includes both copyrighted and public domain material, raising legal concerns regarding reproduction rights under copyright law. Consequently, numerous rightsholders have filed lawsuits against AI developers across various jurisdictions, claiming that the unapproved use of their intellectual property amounts to copyright infringement.[12]

---

[5] Ian Goodfellow, Yoshua Bengio and Aaron Courville, Deep Learning (MIT Press 2016).
[6] KR Chowdhary, Fundamentals of Artificial Intelligence (Springer 2020) 604.
[7] Yufeng G, 'What Is Machine Learning?' (Towards Data Science, 24 August 2017) <https://medium.com/data-science/what-is-machine-learning-8c6871016736> accessed <28 January, 2026>.
[8] Zaid Alissa Almaliki, 'Do You Know How to Choose the Right Machine Learning Algorithm Among 7 Different Types?' (Towards Data Science, 19 March 2019) <https://medium.com/data-science-collective/do-you-know-how-to-choose-the-right-machine-learning-algorithm-among-7-different-types-295d0b0c7f60> accessed <28 January, 2026>.
[9] Ozge Demirci, Jonas Hannane and Xinrong Zhu, 'Who Is AI Replacing? The Impact of Generative AI on Online Freelancing Platforms' (CESifo Working Paper No 11276, 2024).
[10] Jane C Ginsburg, 'The Concept of Authorship in Comparative Copyright Law' (2003) 52 DePaul L Rev 1063, 1066; Amar Nath Sehgal v Union of India [2005] 30 PTC 253 (Del); Copyright Act 1957, s 57 (India).
[11] Mark A Lemley, 'How Generative AI Turns Copyright Upside Down' (2024) 25(2) Columbia Science & Technology Law Review 190.
[12] CISAC, 'Australian Creators Welcome Establishment of Copyright and AI Reference Group' (12 December 2023) <https://www.cisac.org/Newsroom/society-news/australian-creators-welcome-establishment-copyright-and-ai-reference-group> accessed <28 January, 2026>; Mishcon de Reya, 'Generative AI – Intellectual Property Cases and Policy Tracker' (12 August 2024) < https://www.mishcon.com/generative-ai-intellectual-property-cases-and-policy-tracker> accessed <3 February, 2026 >.

II. THE LEGAL FOUNDATIONS: COPYRIGHT AND DATA MINING

Copyright protection has two main objectives which is to encourage the creative efforts of writers and artists and to promote public benefit by ensuring access to creative works.[13] In return for their literary and artistic creations, authors and artists are granted certain exclusive rights.

The right of paternity, for instance, allows authors to claim authorship under their name, a pseudonym, or anonymously, while the right of integrity permits them to resist alterations that might reasonably be seen as damaging to their reputation[14] Copyright holders have the right to take legal action against those they believe have infringed their rights to seek monetary compensation. This includes the possibility of statutory damages, which can vary significantly from $200 to $150,00 for each work based on the severity and intent behind the infringement.[15] Artificial intelligence and copyright law intersect when the generative AI system uses copyright-protected material, absent a license or other defence, and infringes one or more rights of the author. Given the large volume of data, AI models often scrape the publicly available data from the internet en masse.[16] The automated tools "scrape" data from online sources, such as deploying stream ripping software to extract substantial quantities of video or subtitle files from platforms such as YouTube.[17]

Another method entails downloading entire databases, such as comprehensive copies of Wikipedia, via the regularly provided backups offered by the site.[18] A frequently utilized source of training data is text obtained from web crawlers, particularly from resources like Common Crawl.[19] Additionally, developers may incorporate data from third-party sources or non-public repositories.[20] It has been suggested that AI models, having been trained on vast corpora of copyrighted material, may themselves function as derivative works under the Copyright Act, capable, as they are, of reproducing verbatim content from their training data.[21] The output generated by Generative AI violates the author's reproduction rights, as well as the rights to create derivative works based on original material. In the case of *Dow Jones & Co., Inc.* v. *Perplexity AI*, it was highlighted that, depending on the nature of the content and its intended

---

[13] Melville B Nimmer and David Nimmer, Nimmer on Copyright (2021) § 1.03.
[14] Sam Ricketson And Jane C. Ginsburg, International Copyright and Neighboring Rights: The Berne Convention and Beyond, 87, 586 (New York: Oxford University Press, 2nd edn., 2006).
[15] 17 USC § 504.
[16] Brammer v Violent Hues Prods LLC 922 F3d 255 (4th Cir 2019) 266; See, e.g., Anthropic, 'Initial Comments' 5; OpenAI, 'Initial Comments' 5; GenLaw Participants, 'Initial Comments' 44–45.\; IBM Research, Granite Foundation Models (2024) https://www.ibm.com/downloads/cas/X9W4O6BM accessed <28 January, 2026>.
[17] Samantha Cole, 'Leaked Documents Show Nvidia Scraping "A Human Lifetime" of Videos Per Day to Train AI' (404 Media, 5 August 2024) https://www.404media.co/nvidia-ai-scraping-foundational-model-cosmos-project/ accessed <29 January, 2026>; Leo Gao and others, 'The Pile: An 800GB Dataset of Diverse Text for Language Modeling' (2020) 26.
[18] Wikimedia, 'Wikimedia Downloads' https://dumps.wikimedia.org/ accessed <29 January, 2026>; Tom B Brown and others, 'Language Models Are Few-Shot Learners' (2020) 8–9.
[19] Common Crawl, https://commoncrawl.org/ accessed <30 January, 2026>.
[20] Thomas Heldrup, Report on Pirated Content Used in the Training of Generative AI (Rights Alliance 2025) 5–6 https://rettighedsalliancen.dk/wp-content/uploads/2025/03/Report-on-pirated-content-used-in-training-ofAI.pdf accessed <30 January, 2026>; Ashley Belanger, 'Meta Claims Torrenting Pirated Books Isn't Illegal Without Proof of Seeding' (Ars Technica, 20 February 2025) https://arstechnica.com/tech-policy/2025/02/meta-defends-its-vast-book-torrenting-were-just-a-leech-no-proofof-seeding/ accessed <29 January, 2026>.
[21] News/Media Alliance, White Paper: How the Pervasive Copying of Expressive Works to Train and Fuel Generative Artificial Intelligence Systems Is Copyright Infringement and Not a Fair Use (31 October 2023) https://www.newsmediaalliance.org/generative-ai-white-paper/ accessed <31 January, 2026>.

audience, AI systems may also engage issues related to public display and public performance rights.[22] With the growing focus on AI training data, developers have become noticeably more secretive about what their datasets contain, a change that many consider intentional. Consequently, organizations advocating for rightsholders globally are demanding legal requirements for AI developers to disclose information about their training datasets. This transparency is aimed at helping rightsholders assert their rights over their content.[23] Further, to better understand the real-world implications of these issues, we now turn to an illustrative case study of Ghibli Studios. It offers an insight into how opacity in training data and algorithmic design can produce ethically and legally fraught outcomes, reinforcing the need for stronger transparency norms and regulatory oversight.

III. STYLE VS. SUBSTANCE: CASE STUDY OF GHIBLI STUDIOS

Generative AI has considerably altered how creative works are produced, distributed, and consumed. Tools such as ChatGPT, DALL-E, Midjourney, and Leonardo now allow users to generate high-quality imagery that replicates recognizable artistic styles with striking accuracy.[24] Yet many of these models are trained on unlicensed internet data, often including artists' work without their knowledge or consent.[25] This tension came into sharp relief in early 2025, when a viral trend saw users generating images in the distinctive visual style of Studio Ghibli, the celebrated Japanese animation house behind *Spirited Away* and *My Neighbour Totoro*, using OpenAI's image generation tools.[26] The trend attracted significant public attention and placed considerable strain on OpenAI's infrastructure.[27] Beyond the technical disruption, it exposed deeper questions around copyright, artistic identity, and the limits of existing regulatory frameworks, particularly given the prevalence of similar AI-generated aesthetics across platforms such as Instagram and TikTok.[28] At the heart of these concerns lies a structural gap: copyright law, designed for human authors of fixed works, appears ill-equipped to address AI outputs that reproduce artistic style without copying expression directly. [29] Studio Ghibli's films are characterised by hand-drawn animation, soft colour palettes, fluid motion, and recurring

---

[22] Dow Jones & Co Inc v Perplexity AI Inc No 24-cv-7984 (SDNY, 21 October 2024).

[23] Professional Photographers of America, 'PPA's Comments on the Copyright Office NOI on Generative Artificial Intelligence' (17 November 2023) accessed <31 January, 2026>; News/Media Alliance, 'Global Principles on Artificial Intelligence (AI)' (6 September 2023) < https://www.newsmediaalliance.org/global-principles-on-artificial-intelligence-ai/> accessed <31 January, 2026>.

[24] B M Nagapushpa, P Kanchana and P Monish, 'Revolutionizing Art and Design Through AI: Balancing Innovation, Ethics, and Future Preparedness' (2024) International Journal of Scientific Research in Engineering and Management.

[25] W Vivaldi and I Sutedja, 'Using Deep Learning and CBIR to Address Copyright Concerns of AI-Generated Art: A Systematic Literature Review' (2024) 5(10) Devotion.

[26] Tulisha Srivastava, 'Spirited Away, Totoro & More Unite in New Official Studio Ghibli Design for TOHO Movie Theater Campaign' (CBR, 12 April 2026) https://www.cbr.com/studio-ghibli-new-design-toho-movie-theater-campaign/ accessed 15 April 2026.

[27] James Vincent, 'OpenAI Says "Our GPUs Are Melting" as It Limits ChatGPT Image Generation Requests' (The Verge, 28 March 2025) https://www.theverge.com/news/637542/chatgpt-says-our-gpus-are-melting-as-it-puts-limit-on-image-generation-requests accessed < February 19, 2026 >.

[28] Adil S Al-Busaidi and others, 'Redefining Boundaries in Innovation and Knowledge Domains: Investigating the Impact of Generative Artificial Intelligence on Copyright and Intellectual Property Rights' (2024) 9(4) Journal of Innovation & Knowledge 100630 https://doi.org/10.1016/j.jik.2024.100630 accessed < February 19, 2026 >.

[29] W Vivaldi and I Sutedja, 'Using Deep Learning and CBIR to Address Copyright Concerns of AI-Generated Art: A Systematic Literature Review' (2024); Chris Stokel-Walker, 'Studio Ghibli and the Legal Gray Zone of AI-Generated Art' (Business Insider, 29 March 2025) accessed <31 January, 2026>.

thematic preoccupations, among them childhood innocence, environmental stewardship, and quiet resilience, that together lend their works a visual and narrative distinctiveness widely associated with the studio.[30] While the phrase "Ghibli aesthetic" has gained currency in popular discourse, it is worth noting that artistic style, as such, does not attract copyright protection; what the law may protect, if anything, is the specific expressive elements through which that style is realised.[31] Advances in generative AI, particularly in diffusion and transformer-based models such as DALL-E, Midjourney, and Leonardo.AI, allow users to create original images in the Ghibli style by submitting simple text prompts.[32] These models are developed using extensive datasets filled with artistic images, allowing them to recognize the patterns, textures, and emotional signals that define particular styles. Tools like the Ghibli Diffusion model are specifically designed to replicate Studio Ghibli's artistic features, including pastel brushwork, watercolours, and childlike facial expressions.[33] Culturally, the replication of the Ghibli aesthetic speaks to the nostalgic allure and emotional impact of the studio's work. The incorporation of Ghibli-inspired elements by AI image generators brings up questions about attribution, since these systems typically do not produce original creative work. The main problem stems from the outdated criteria for determining authorship according to current legal standards. The U.S. Copyright Office enforces copyright protection only for human-authored works according to their current established policy.[34] Therefore, currently, AI-generated images remain unprotected by conventional laws because courts recognize them as machine-produced rather than human works; thus, there is an unprotected space with ownership challenges. Under established copyright doctrine, protection attaches to specific fixed expressions rather than to style itself, meaning that visual characteristics such as color palette, line quality, or compositional mood remain, in principle, freely replicable.[35] Applied to the Ghibli context, this suggests that AI-generated imagery evoking the studio's recognizable visual language may not constitute infringement, provided no particular protected work is directly reproduced.[36] Whether this outcome reflects sound policy is another matter; critics argue that the doctrine, however internally coherent, leaves studios like Ghibli practically defenseless against systematic stylistic appropriation at scale. The legal permissibility of style imitation does not fully explain why Ghibli-style AI imagery spread as rapidly as it did. Several platform-level dynamics appear to have accelerated its circulation. Algorithmic recommendation systems on Instagram and TikTok tend to prioritise visually distinctive content, and Ghibli-derived imagery, with its soft palettes and immediately legible aesthetic,

[30] MotionCue, '5 Animation Lessons from Studio Ghibli' < https://motioncue.com/animation-lessons-from-studio-ghibli/> accessed <30 January, 2026>.
[31] Nichols v. Universal Pictures Corp., 45 F.2d 119 (2d Cir. 1930).
[32] Xinyue Fan and Yifan Liang, 'The Research on the Characteristics of AI Application in Art Field and Its Value' in Proceedings of the International Conference on Advances in Computer Science and Communication Technology.
[33] Getimg.ai, 'Studio Ghibli Free AI Generator' <https://getimg.ai/models/ghibli-diffusion> accessed <20 January 2025>.
[34] US Copyright Office, Copyright and Artificial Intelligence, Part 2: Copyrightability < https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-2-Copyrightability-Report.pdf > accessed <28 January 2025>.
[35] Ibid.
[36] W Vivaldi and I Sutedja, 'Using Deep Learning and CBIR to Address Copyright Concerns of AI-Generated Art: A Systematic Literature Review' (2024) 5(10) Devotion.

lends itself well to this kind of amplification.[37] Equally significant is the ease of generation itself: users require no artistic training to produce studio-quality imagery, lowering the barrier to participation and broadening the pool of contributors. Meme-like replicability, combined with the nostalgic resonance of the Ghibli visual language, likely reinforced sharing behavior, images of this kind activate something recognizable and emotionally familiar, which platforms reward with reach.[38] The result is a self-sustaining content cycle, in which visibility drives participation, and participation drives further visibility. Viral content trends carry consequences that extend well beyond the social. The computational demands of generative AI, spanning both training and inference, are considerable, and a surge in image generation requests can deplete GPU capacity rapidly, degrading system performance and increasing energy expenditure.[39] The 2025 Ghibli trend offered a concrete illustration: OpenAI publicly acknowledged the strain on its infrastructure as requests spiked. It is worth pausing on what this implies. Viral AI content is not merely a cultural phenomenon; it is a resource-intensive event, with environmental and infrastructural costs that tend to go unacknowledged in public discourse around creative AI.[40] Returning to the legal dimension, the absence of protection for style leaves artists in a structurally weak position. Training datasets routinely incorporate artists' work without consent, and the outputs those datasets enable can replicate a creator's distinctive visual identity without triggering any enforceable claim. For Studio Ghibli, a studio whose aesthetic is among the most recognizable in contemporary animation, this gap is particularly acute. Existing frameworks were not designed with generative AI in mind, and their application to these circumstances remains unsettled, a point that courts and legislators across jurisdictions are only beginning to confront.[41]

IV. LEGAL UNCERTAINTIES: WHO IS THE AUTHOR?

A. The Human-Centric Premise of Copyright

The reverberations of the AI applications to create literary and artistic works are palpable in the ongoing discourse surrounding the intricate matter of 'authorship' of the works generated by AI.[42] Traditionally, the notion of authorship has sought to promote and protect the uniqueness of human writers. The Berne Convention, a foundational international convention for copyright protection, assures that literary and creative works are protected worldwide, with advantages extended to writers and their successors in title.[43] However, this protection is not necessarily extendable to AI generated works, as even though

[37] Dinesh Deckker and Subhashini Sumanasekara, 'Dreams and Data: Ghibli-Style Art, Copyright, and the Rise of Viral AI Imagery' (TechRxiv, 1 April 2025) <https://www.techrxiv.org/doi/full/10.36227/techrxiv.174353158.87082654/v1> accessed <28 January 2025 >.
[38] Xinyue Fan and Yifan Liang, 'The Research on the Characteristics of AI Application in Art Field and Its Value' in Proceedings of the International Conference on Advances in Computer Science and Communication Technology (2023).
[39] Supra note 26.
[40] Trebor Scholz (ed), Digital Labor: The Internet as Playground and Factory (Routledge 2013).
[41] Peng Xiang and others, 'The Impact of the COVID-19 Pandemic on Nostalgic Social Media Use' (2024) 15 Frontiers in Psychology art 1431184.
[42] Berne Convention for the Protection of Literary and Artistic Works (4 May 1896, entered into force 29 January 1970) 828 UNTS 221.
[43] Gaffar H and S Albarashdi, 'Copyright Protection for AI-Generated Works: Exploring Originality and Ownership in a Digital Landscape' (2025) 15(1) Asian Journal of International Law 23, 31.

the Berne Convention does not clearly define the word "author", the assumption to include the AI-generated works in the realm of authorship will be detrimental.[44] The Convention shadows a humanist nature, and granting copyright to the AI-generated work will lose the soul of copyright. AI should be viewed as a tool used by the human author, rather than an autonomous creator in its own right.[45] The *Feist Publications, Inc. v. Rural Telephone Service Company* case has affirmed that an "author" refers to a human writer and composer of an original creative work. The work must showcase a 'modicum of creativity' and originality traceable to the author. It indicates that factual information or data aggregation without the requisite creative input does not fulfil the threshold for copyright protection.[46] The case of *University of London Press Ltd v University Tutorial Press Ltd* has defined originality as, "an expression of ideas in the original form, i.e., it originates from the author in the sense that the work must not be a copy of another work."[47] Moreover, Section 9(3) of the UK Copyright, Designs and Patents Act (CDPA) of 1988 stipulates that the authorship of the person who makes the necessary arrangements for the creation of computer-generated work should be regarded under copyright protection.[48] There is a possibility that 'necessary arrangement' refers to the original undertakings of the author who created the copyrighted work, used as an input in the training process of AI. In *Thaler* v. *Perlmutter*, the court ruled that content generated by AI, lacking human creation, cannot receive copyright protection due to the requirement that legal authorship belong to natural persons. [49] The technical nature of human input, combined with the automated processes of AI algorithms, which are still evolving and lack clarity, does not sufficiently justify copyright protection for AI-generated works.[50] This perspective is also reflected in the legislation of other countries such as New Zealand,[51] Hong Kong,[52] Ireland,[53] and China.[54] So this fills a gap in defining the term "author", and continues the tradition of prohibiting nonhuman authors from holding copyright in order to promote the social purpose of the Copyright Law. The Copyright Law also reflects certain exceptions, which are encoded in the US Copyright clause. The interpretation of Title 17 of the U.S. Code has consistently maintained the stance that copyright protection is limited to works created by human authors.[55] The title 17 USC § 107 of the 1976 Copyright Act also incorporates a "*fair use doctrine*" to determine copyright infringement of the new work.[56] In *Harper & Row Publishers v. Nation Enterprises*, the court observed that the fair use should not be interpreted broadly as to swallow the commercial value of a copyrightable work.[57] It must ensure that

[44] Paul Goldstein, International Copyright: Principles, Law, and Practice (OUP 2001).
[45] Sam Ricketson and Jane C Ginsburg, International Copyright and Neighboring Rights: The Berne Convention and Beyond (2nd edn, vol 1, OUP 2005) 358.
[46] Feist Publications Inc v Rural Telephone Service Co 499 US 340 (1991).
[47] University of London Press Ltd v University Tutorial Press Ltd [1916] 2 Ch 601.
[48] Copyright, Designs and Patents Act 1988, s 9(3).
[49] Thaler v Perlmutter 130 F4th 1039 (DC Cir 2025).
[50] WIPO Secretariat, 'Revised Issue Paper on Intellectual Property Policy and Artificial Intelligence' (WIPO/IP/AI/2/GE/20/1 Rev, 11 June 2022) para 23.
[51] Copyright Act 1994 (NZ), s 5(1)–(2)(a).
[52] Copyright Ordinance (Cap 528) (HK), s 11(3).
[53] Copyright and Related Rights Act 2000 (Ireland), s 21(f).
[54] China IPR SME Helpdesk, Copyright Protection in China: A Guide for European SMEs (European Commission 2010).
[55] 17 USC § 107.
[56] Ibid.
[57] Harper & Row Publishers Inc v Nation Enterprises 471 US 539, 562–63 (1985).

the new work has a distinct purpose and character and is not merely transformative in character. Generative AI focuses on capturing the specific expressive choices made by the author, such as the words an author used rather than their abstract ideas, and this can raise legal concerns, as it may involve copying protected expression, potentially violating the reproduction right and, in some cases, the derivative works right..[58] The 2025 judgment in *Thomson Reuters Enterprise Centre GmbH v. Ross Intelligence Inc.* significantly undermined the "fair use" defence frequently invoked by AI companies. It reinforced that training AI on publicly accessible, commercially licensed content that embodies intellectual labour may constitute copyright infringement. This results in a shift from the era of unchecked web scraping to an age of accountability for the AI companies.[59] The United States Copyright Office recently cancelled the copyright protection of the AI-generated images within the graphic novel "*Zarya of the Dawn*". The copyright protection was rejected because the images were not a product of human authorship.[60] In India, the case of *Amar Nath Sehgal v. Union of India* advances Section 57 of the Copyright Act, 1957, stating that the thoughts and sentiments of the human author, and not AI, influence moral rights. There is no exception under copyright law for text and data mining without obtaining a license from copyright owners.[61] The usage of the "sweat of the brow doctrine," which recognized that the labor of the author is sufficient for granting copyright protection, has been discarded. It has instead, adopted the approach of "modicum of creativity" from 2008, which accepts a work to be to be original if there is some substantive and subjective variation.[62] The work created by AI is an objective output that is based on the subjective qualifications and the prompts of the authors, who should be thereby granted a copyright protection.[63] It is advanced by Locke's labor theory as a justification to copyright, emphasizing the "individual's right to the fruits of their labor." The creator of a text prompt puts their creativity, much like effort, into developing the prompt. As a result, they ought to have rights to copyright protection for the resulting output generated by AI.[64] Therefore, an analysis of copyright protection of AI-generated work and its ownership involves the examination of the element of originality and the identification of the author of the work. As mentioned in the preceding paragraphs, AI tools generate work by relying on data that they have been trained on. They are trained on the data which is already available on the internet and the pre-existing data which may be copyrighted. The fact that an individual can claim copyright over the originality of the work used as an input in the AI-generated work does not automatically confer protection over the AI output. Hence, the copyright law should be protected by defining the term

[58] Authors Guild v Google Inc 804 F3d 202 (2d Cir 2015).
[59] Congressional Research Service, Generative Artificial Intelligence and Copyright Law (2026) https://www.congress.gov/crs-product/LSB10922 accessed <February 18, 2026 >.
[60] Barton Beebe, 'An Empirical Study of US Copyright Fair Use Opinions, 1978–2005' (2008) 156 University of Pennsylvania Law Review 549, 605 https://scholarship.law.upenn.edu/penn_law_review/vol156/iss3/1 accessed < February 18, 2026 >.
[61] Amar Nath Sehgal v Union of India [2005] 30 PTC 253 (Del).
[62] Alexander D Northover, 'Enough and as Good in the Intellectual Commons: A Lockean Theory of Copyright and the Merger Doctrine' (2016) 65 Emory Law Journal 1363, 1376.
[63] Darin Glasser, 'Copyrights in Computer-Generated Works: Whom, if Anyone, Do We Reward?' (2001) 1 Duke Law & Technology Review 1.
[64] Emmanuel Salami, 'AI-Generated Works and Copyright Law: Towards a Union of Strange Bedfellows' (2021) 16(2) Journal of Intellectual Property Law & Practice 124.

"authorship" with the originality and creativity of human authors, who should be accorded protection.

Legal Gaps and Policy Dilemmas in Attributing Authorship to Artificial Intelligence As a starting point, the vast majority of the material used to train most generative AI models is typically safeguarded by copyright law.[65] The content on Wikipedia is protected by copyright and is available under Creative Commons (CC) or similar 'copyleft' licenses, which may allow for its use in training generative AI models. The CC License provides an identification of the creator(s) of the licensed material and permits the use of copyrighted material. However, since these licenses provide access to only a limited portion of copyright-protected content, it's improbable that state-of-the-art generative AI models could be developed solely with CC-licensed material.[66] This creates an infringement of the copyright law in the training process of AI content. A report by the Washington Post in April 2024 revealed that numerous companies engaged in AI development avoid maintaining internal records of their training data due to concerns that this information could serve as proof of copyright violations or breaches of data protection regulations.[67] Taking the example of OpenAI, it has shifted from the detailed explanations of the training data used to train a particular model to single-sentence descriptions. The method that offers the greatest level of transparency is for AI developers to make the training datasets for a specific AI model completely open to the public. With this 'full access' strategy, third parties, including rights-holders, can examine the training data directly and verify its content independently. Additionally, enhancing transparency can foster public trust in AI models by narrowing the information gap between model creators and users, thus preventing a 'market for lemons' situation where the demand for generative AI models diminishes.[68] The recent working paper on Generative AI and Copyright, released by the Department for Promotion of Industry and Internal Trade (DPIIT), has anchored a "one nation, one license, one payment" principle to collect royalties from AI systems trained on copyrighted content. A government-appointed committee will fix the royalties and distribute them to the right holders. Moreover, Section 2(d)(vi) of the Indian Copyright Act has an important significance towards human creation of work.[69] It provides an attribution to the creative human input that directly contributes to the output of the work generated by AI tools like ChatGPT, Google Bard, DeepAI, etc. The developer/programmer who programs and trains the AI can be considered to have "caused the work to be created" and would have a better claim than the user under Section 2(d)(vi).[70] A further comparison is also made between the way AI operates and human creativity. It is believed that

[65] Paris Act Relating to the Berne Convention for the Protection of Literary and Artistic Works (9 September 1886, entered into force 15 December 1972) 3 UNTS 1.
[66] Aaron Gokaslan and others, 'Common Canvas: Open Diffusion Models Trained on Creative-Commons Images' in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024) 8250.
[67] US Copyright Office, Copyright and Artificial Intelligence, Part 1: Digital Replicas (2024) 53–56 <https://www.copyright.gov/ai/Copyright-and-Artificial-Intelligence-Part-1-Digital-Replicas-Report.pdf> accessed < February 19, 2026 >.
[68] George A Akerlof, 'The Market for "Lemons": Quality Uncertainty and the Market Mechanism' (1970) 84 Quarterly Journal of Economics 488.
[69] Copyright Act 1957 (India), s 2(d).
[70] Naithani P, 'Issues of Authorship and Ownership in Work Created by Artificial Intelligence – Indian Copyright Law Perspective' (2022) 11(1) NTUT Journal of Intellectual Property Law & Management 7.

humans are creative, whereas AI relies on human creativity for the generation of work. The court in India, through the case of *Rupendra Kashyap v. Jiwan Publishing House Pvt. Ltd.* has presented that AI does not require any incentive to generate work, unlike humans.[71] Therefore, the motivation and stimulation that humans seek through copyright protection and authorial rights to undertake various innovations do not apply to AI.[72] It reserves the concept of authorship to humans and not only affirms basic human values but also stands as a welcome reminder of human individuality and uniqueness. The Copyright Office or the judiciary has not yet claimed AI tools as an author, and humans have also not attempted to register copyright for the AI-generated works.[73] In India, though there is a rumour that the Copyright Office has registered an AI-generated poem under the name of "AI-Dada", however, there is still no confirmation.[74] Therefore, the granting of authorship rights to the AI-generated works does not seem to be possible in the near future. The copyright law still recognizes the concept of authorial rights to humans in AI-generated works if they are original, as there still exists human involvement in the work creation process.[75]

B. Evolving Notions of Originality and Copyrighted Content

Countries like the UK, Hong Kong, Ireland, New Zealand, and India have made meaningful progress in tackling the challenges associated with AI-generated works. These nations have proactively revised their laws to recognize the copyrightability of works that incorporate both computer-generated elements and human contributions.[76] In doing so, they have taken a more advanced stance than the international legal community, offering clearer legal guidance in this field.[77] Despite this, the existing laws do not fully tackle situations where a computer produces a work with only minimal or indirect human participation, which may create a gap in the legal system.[78] In comparison, many jurisdictions around the globe continue to find it challenging to confront the intricate issue of authorship and often remain mute on this controversial topic.[79] As AI becomes an increasingly active participant in the creative process, the question of who or what qualifies as an "author" demands reconsideration. There is a broadening of the term "author" based on ideas of labour, personality, and communication,

---

[71] Rupendra Kashyap v Jiwan Publishing House 1996 (38) DRJ 81.
[72] Gerald Dworkin, 'The Moral Right of the Author: Moral Rights and the Common Law Countries' (1994–95) 19 Columbia-VLA Journal of Law & the Arts 229.
[73] US Copyright Office, Copyright Protections for Press Publishers (2022) https://www.copyright.gov/policy/publishersprotections/202206-Publishers-ProtectionsStudy.pdf accessed <3 February 2026 >.
[74] Ibid.
[75] Michel A, 'AI-Generated Creations: Challenging the Traditional Concept of Copyright, a Research into the Question of Whether Works Created by an Artificial Intelligence Program Have Copyright Protection in the Netherlands and the European Union' (Thesis, Tilburg University 2018) 27–28 http://arno.uvt.nl/show.cgi?fid=148002 accessed < 3 February 2026 >.
[76] Nattapong Suwan-In, 'Copyright Protection on AI-Generated Work: The Case Study of the US, UK, and Thailand Copyright Laws' (2021) Journal of Law, Public Administration and Social Science 131, 142; Copyright Act 1968 (Cth); IceTV Pty Ltd v Nine Network Australia Pty Ltd [2009] HCA 14.
[77] Copyright, Designs and Patents Act 1988 (UK), s 9(3); Copyright Act 1994 (NZ), s 5; Copyright Ordinance (Cap 528) (HK), s 11(3); Copyright and Related Rights Act 2000 (Ireland), s 21(f).
[78] Comparative Law Library Unit, European Parliamentary Research Service, Copyright Law in the EU: Salient Features of Copyright Law Across the EU Member States (European Parliament 2018) https://www.europarl.europa.eu/RegData/etudes/STUD/2018/625126/EPRS_STU(2018)625126_EN.pdf accessed < 5 February 2026 >.
[79] Daniel Bötticher, 'Copyright Protection for Artificial Intelligence Output' (Master's thesis, University of St Gallen 2019) 33–35.

which is suggested by Burrow Giles Court.[80] This formula is an evolution, as opposed to the death of the copyright author. The judgment of *Urantia Foundation v. Maaherra*, which has elaborated upon the element of human creativity, has been moulded to include AI authorship.[81] It has been stated that the end user communicates to the AI, which infuses its labour, communication, and personality to create the work.[82] It balances and recognizes the authorship of AI along with human originality and creativity. The Berne Convention has outlineda 'three-step test' indicating that any exceptions to the reproduction of copyrighted material must serve a specific purpose, must not interfere with the usual exploitation of the work, and must not unjustly harm the legitimate interests of the author.[83] This implies that exceptions allowing the unauthorized use of copyrighted content for training generative AI models should not be intended for commercial purposes.[84] Artist Rembrandt van Rijn's recent project Rembrandt, has aimed to create a completely new art piece through AI generation, .[85],seeking to acknowledge the copyrightability of artificial intelligence by attributing a certain level of personality to it. Additionally, it emphasizes that human contributions to the creation and processing of such works will receive appropriate recognition and copyright ownership.[86] In June 2023, the European Parliament approved a measure that would mandate generative AI providers to create and publicly share a detailed summary of their use of training data that is protected by copyright..[87] The critique of this approach was further incorporated into Article 53(1)(c), which mandates that providers of general-purpose AI models deliver a summary that is broadly comprehensive rather than overly technical.[88] This aims to assist parties with legitimate interests, including copyright holders, in exercising and enforcing their rights under Union law. For instance, it should include a listing of the primary data collections or datasets utilized for training the model, such as significant private or public databases or data archives, along with a narrative explanation detailing other data sources employed.[89] Even with this guidance, there is still significant uncertainty regarding what is expected from GPAI model providers. Although AI is not yet recognized as an independent legal author in most legal systems, its functional contribution is increasingly being taken into account in legislative reforms and judicial reasoning. This indicates an emerging willingness to consider AI-related interests within the copyright framework, even as human creativity remains the central normative foundation.

---

[80] Burrow-Giles Lithographic Co v Sarony 111 US 53 (1884).
[81] Urantia Foundation v Maaherra 114 F3d 955, 964 (9th Cir 1997).
[82] Andres Guadamuz, 'Artificial Intelligence and Copyright' (WIPO Magazine, 2017) https://www.wipo.int/wipo_magazine/en/2017/05/article_0003.html accessed < 7 February 2026 >.
[83] Berne Convention for the Protection of Literary and Artistic Works (9 September 1886) art 9(2); Eleonora Rosati, 'No Step-free Copyright Exceptions: The Role of the Three-step Test in Defining Permitted Uses of Protected Content (Including TDM for AI-Training Purposes)' (2024) 46 European Intellectual Property Review 262.
[84] Saliltorn Thongmeensuk, 'Rethinking Copyright Exceptions in the Era of Generative AI: Balancing Innovation and Intellectual Property Protection' (2024) The Journal of World Intellectual Property 278, 284.
[85] Péter Mezei, 'From Leonardo to the Next Rembrandt – The Need for AI-Pessimism in the Age of Algorithms' (4 May 2020) 7 https://ssrn.com/abstract=3592187.
[86] Pamela Samuelson, 'AI Authorship?' (2020) 63(7) Communications of the ACM 22.
[87] European Parliament, Amendments Adopted on 14 June 2023 on the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) art 28b(4)(a).
[88] Ibid.
[89] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) (14 June 2023) recital 107.

V. TOWARD A FAIRER FUTURE: TRANSPARENCY AND REFORM

AI and copyright laws are a strange mix, and there is a need to figure out how to deal with these tricky issues and make rules about them. This article shows that while training data and transparency requirements have some good points, they don't work the same way in different countries because of their own copyright laws. These rules alone can't solve the problem with using copyrighted stuff to train generative AI models. Transparency is important for accountability and public trust in AI training data. There is a need for laws that force AI developers to tell us what kind of data they use, how much, and where it came from. But we also need to come up with new ways to do this that are more effective. The current laws don't address the main flaws in how we use copyrighted content with AI. Some countries have copyright laws that say the person who created the original work is the author. But these laws weren't made with generative AI in mind. The transparency rules in the AI Act are unlikely to help individual authors much. At present, we must determine the appropriate strategies to modify copyright laws in order to harmonize the interests of all stakeholders impacted by generative AI.

There is no universally applicable solution, as the most effective approach will vary depending on the unique legal frameworks, economic conditions, and cultural norms of each nation. Policymakers worldwide can team up with key stakeholders to figure out the best policies for their specific situations. With AI developing quickly, these policies need to be checked regularly to make sure they are still relevant and effective. This is all while dealing with the sometimes-overblown hype about AI's economic potential and the growing backlash against AI technology from the public, especially those in creative industries. Artists are at risk of having their unique artistic styles copied by AI systems without their permission, especially when those styles are used to make derivative works that look and feel almost exactly like the original.Policymakers should propose new laws that provide clear and specific protections for unique, identifiable artistic styles.

This means figuring out what counts as originality and coming up with other ways to account for the special nature of these works. International cooperation is crucial for setting consistent standards and guidelines across different countries that make sure everything works well, allows for cross-border recognition, and protects AI-generated works. Standardized licensing and attribution systems for AI-generated works can be created to make it easy to recognize, own, and give credit to human creators and AI systems. Finally, raising public awareness through education initiatives is crucial for helping creators, users, and the general public grasp the nuances of AI-generated content, including its implications as well as their rights and responsibilities. By focusing on these recommendations, policymakers and researchers can establish a flexible and supportive framework that fosters innovation, safeguards rights, and encourages the responsible application of AI in creative pursuits.